

## Regressionsanalyse bei fehlenden Variablenwerten (missing values): Imputation oder Nicht-Imputation? Eine Anleitung für die Regressionspraxis mit SPSS

Urban, Dieter; Mayerl, Jochen; Wahl, Andreas

Veröffentlichungsversion / Published Version

Arbeitspapier / working paper

### Empfohlene Zitierung / Suggested Citation:

Urban, D., Mayerl, J., & Wahl, A. (2016). *Regressionsanalyse bei fehlenden Variablenwerten (missing values): Imputation oder Nicht-Imputation? Eine Anleitung für die Regressionspraxis mit SPSS*. (2. korrigierte Aufl.) (Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart -SISS-, 44). Stuttgart: Universität Stuttgart, Fak. 10 Wirtschafts- und Sozialwissenschaften, Institut für Sozialwissenschaften. <https://doi.org/10.18419/opus-9112>

### Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

**gesis**  
Leibniz-Institut  
für Sozialwissenschaften

### Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Mitglied der  
  
Leibniz-Gemeinschaft

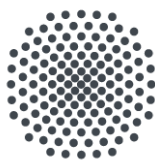
**SISS:**  
**Schriftenreihe des Instituts für Sozialwissenschaften  
der Universität Stuttgart**

**No. 44 / 2016**

***Regressionsanalyse bei fehlenden  
Variablenwerten (missing values):  
Imputation oder Nicht-Imputation?  
Eine Anleitung für die Regressionspraxis  
mit SPSS.***

***(2. korrigierte Auflage)***

***Dieter Urban  
Jochen Mayerl  
Andreas Wahl***



**Universität Stuttgart**

**Abteilung für Soziologie  
und empirische Sozialforschung**



**SISS:**  
**Schriftenreihe**  
**des Instituts für Sozialwissenschaften**  
**der Universität Stuttgart:**  
**No. 44 / 2016**

Regressionsanalyse bei fehlenden Variablenwerten  
(missing values): Imputation oder Nicht-Imputation?  
Eine Anleitung für die Regressionspraxis mit SPSS  
(2. korrigierte Auflage)

Dieter Urban  
Jochen Mayerl  
Andreas Wahl

**ISSN 2199-7780**

**Universität Stuttgart**  
**Institut für Sozialwissenschaften**  
**Abt. f. Soziologie u. empirische Sozialforschung**  
**70174 Stuttgart**



**Dieter Urban, Jochen Mayerl und Andreas Wahl**  
**Regressionsanalyse bei fehlenden Variablenwerten (missing values): Imputation oder Nicht-Imputation? Eine Anleitung für die Regressionspraxis mit SPSS.**  
**(2. korrigierte Auflage)**

SISS – Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart. No. 44/2016  
(<http://www.uni-stuttgart.de/soz/institut/forschung/SISS44.pdf>)

**Z U S A M M E N F A S S U N G:** Dieser SISS-Beitrag zeigt, wie mit SPSS eine multiple Regressionsanalyse durchgeführt werden kann, wenn einzelne Variablen des analysierten Datensatzes fehlende Werte (missing values) aufweisen. Es wird erläutert, in welcher Weise das Ausmaß und die Relevanz fehlender Werte ermittelt werden sollte, wann ein listenweiser Ausschluss von Fällen mit fehlenden Werten in einer Regressionsanalyse möglich ist, und unter welchen Bedingungen und in welcher Weise (mit welchen Vor- und Nachteilen) in SPSS das Verfahren der stochastischen Regressionsimputation (sRI) oder der Multiplen Imputation (MI) einzusetzen ist. Die Darstellung ist praxisorientiert. Der Text enthält für alle beschriebenen statistischen Verfahren die entsprechenden SPSS-Steueranweisungen (SPSS-Syntaxfiles).

**Dieter Urban, Jochen Mayerl and Andreas Wahl**  
**Regression analysis when variables have missing values: Imputation or no imputation? A guide for practical regression analysis with SPSS.**  
**(2nd revised edition)**

SISS–Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart. No. 44/2016.  
(<http://www.uni-stuttgart.de/soz/institut/forschung/SISS44.pdf>)

**A B S T R A C T:** This SISS-paper shows, how to perform a multiple regression analysis with SPSS when variables of the concerning data file have missing values. The paper informs about detecting the amount and relevancy of missing values, shows when it is possible to perform a listwise deletion of cases in regression analysis, and describes under which conditions and how to apply a stochastic regression imputation (sRI) or a multiple imputation procedure (MI) in SPSS discussing the advantages and disadvantages of both methods. The argumentation is practice-oriented. The paper contains the appropriate SPSS-syntax files for all statistical techniques described.



## Vorwort zur 2. Auflage

In der zweiten Auflage dieses Textes wurden einige inhaltliche Fehler der ersten Auflage korrigiert und zwei zusätzliche SPSS-Syntaxfiles (12.2 und 12.3) aufgenommen (alle folgenden Seitenangaben beziehen sich auf die 1. Auflage):

Korrektur, S. 24, letzter Absatz

alt: "... werden dazu alle anderen metrischen Modellvariablen benutzt."

neu: "... werden dazu alle anderen Modellvariablen benutzt."

Korrektur, S. 25, erster Absatz

alt: "... Interaktionsvariablen (als Produkt aus metrischen Variablen) verwendet ..."

neu: "... Interaktionsvariablen verwendet ..."

Korrektur, S. 48, SPSS-Syntax 16

alt: "DATASET ACTIVATE *AS\_imputiert* WINDOW=ASIS."

"SORT CASES BY *Imputation\_*."

neu: "DESCRIPTIVES *AS\_Index Alter Bildung*"

"/SAVE."

Modifikation, S. 38, 42, 47

alt: "SPSS-Syntax 12"

neu: "SPSS-Syntax 12.1"

Ergänzung, S. 38, mittig

neu: "Mit SPSS-Syntax 12.1 ... benutzt werden."

Ergänzung, S. 38, unten

neu: SPSS-Syntax 12.2 und SPSS-Syntax 12.3





## 0 Kommentierte Gliederung

In diesem SISS-Beitrag wird die Problematik von fehlenden Werten für die Durchführung von Regressionsanalysen beschrieben. Es werden verschiedene Strategien zum Umgang mit fehlenden Werten in der Regressionsanalyse vorgestellt, kritisch diskutiert sowie praktische Anweisungen für die Durchführung von Regressionsanalysen mit fehlenden Werten unter Verwendung von SPSS gegeben. Im Einzelnen behandeln die Kapitel folgende Themen:

- Kapitel 1: Es wird die Problematik fehlender Werte für die Schätzung von Regressionsmodellen erläutert und es wird ein Überblick über die Strategien gegeben, die SPSS zum Umgang mit fehlenden Werten in der Regressionsanalyse anbietet.
- Kapitel 2: Es wird gezeigt, in welcher Weise mit SPSS zu untersuchen ist, ob ein Datensatz eine problematische Anzahl von fehlenden Werten aufweist, und welche Ausfallmechanismen für den Umgang mit fehlenden Werten zu unterscheiden sind.
- Kapitel 3.1: Es wird erläutert, unter welchen Bedingungen ein listenweiser Ausschluss von Fällen mit fehlenden Werten in einer Regressionsanalyse inhaltlich zu rechtfertigen ist und wie der listenweise Ausschluss von Fällen mit SPSS-Syntax durchzuführen ist.
- Kapitel 3.2: Es wird die stochastische Regressionsimputation (sRI) als Verfahren zur statistischen Schätzung von fehlenden Werten und zur Imputation entsprechender Schätzwerte in einen SPSS-Datensatz erläutert sowie die praktische Durchführung dieses Verfahrens mit SPSS-Syntax vorgestellt. Zudem werden die Vor- und Nachteile dieses Verfahrens benannt.
- Kapitel 3.3: Es wird das Verfahren der Multiplen Imputation (MI) als Alternative zur stochastischen Regressionsimputation (sRI) erläutert und die praktische Durchführung dieses Verfahrens mit SPSS-Syntax vorgestellt. Zudem werden die Vor- und Nachteile dieses Verfahrens benannt.
- Kapitel 4: Es wird vor der besonderen Problematik, die bei Durchführung von Imputationsverfahren mit SPSS gegeben ist, gewarnt und es wird gezeigt, durch welche einfache Analysestrategie diese Problematik zu entschärfen ist.
- Anhang: Es wird die FCS-Methode (Fully Conditional Specification Methode), die wir bei Durchführung einer multiplen Imputation mit SPSS empfehlen, auf leicht verständliche Weise (ohne Bayes-Statistik) erläutert.

## 1 Die Problematik fehlender Werte in der Regressionsanalyse

Die Daten für eine Regressionsanalyse sind oftmals unvollständig. So können z.B. einzelne, für eine Befragung ausgewählte Personen die Teilnahme an einer Studie komplett verweigern. Oder bestimmte Unternehmen können bei einer vergleichenden Organisationsstudie ihre betriebsbezogenen Daten aus Geheimhaltungsgründen nicht weitergeben wollen. In diesen Fällen spricht die Sozialforschung von fehlenden Fällen infolge eines „unit nonresponse“.

Personen oder Organisationen können sich aber auch an einer Befragung beteiligen und nur die Antworten zu einzelnen Fragen (z.B. zu Fragen hinsichtlich ihres Einkommen oder ihrer Gewinnkalkulationen) verweigern. Als Folge eines solchen „item nonresponse“ entstehen fehlende Werte, die in der Sozialforschung auch als „missings“ oder „missing values“ bezeichnet werden.

Zu unterscheiden sind also „unit nonresponses“, die zu „fehlenden Fällen“ führen, und „item nonresponses“, die zu „fehlenden Werten“ führen.

Im Folgenden werden wir uns nur mit Strategien zum Umgang mit fehlenden Werten und nicht mit fehlenden Fällen beschäftigen. Zwar lässt sich auch auf systematisch fehlende Fälle mit verschiedenen Gewichtungsverfahren reagieren. Jedoch sind solche Gewichtungen nur schwierig zu entwickeln und in ihrer Anwendung häufig sehr risikoreich, sodass wir sie nicht in diesem SISS-Beitrag behandeln wollen.<sup>1</sup>

Bei Datensätzen mit fehlenden Werten muss darüber entschieden werden, wie mit denjenigen Untersuchungsfällen, die bei einigen Frage- bzw. Beobachtungselementen fehlende Werte aufweisen, in der statistischen Analyse umzugehen ist. Zwar können unter gewissen Bedingungen (dazu mehr im folgenden SISS-Kapitel 3.1) fehlende Werte für die Ergebnisse einer Regressionsanalyse bedeutungslos sein, sodass sie ignoriert werden können. Oftmals jedoch können fehlende Werte die Ergebnisse einer Regressionsschätzung auch ganz wesentlich verfälschen. Insbesondere können sie:

- a) die Schätzwerte der Regressionskoeffizienten verzerren,
- b) die Schätzwerte der Standardfehler von Regressionskoeffizienten verzerren,
- c) die Ergebnisse von Signifikanztests verfälschen,
- d) die Teststärke von Regressionsanalysen beeinflussen.<sup>2</sup>

---

<sup>1</sup> Vgl. dazu Gabler/Ganninger 2010.

<sup>2</sup> Zur Bedeutung der Teststärke in Regressionsanalysen vgl. Urban/Mayerl 2011: 156ff.

Wann sind also fehlende Werte harmlos und wann können sie die Ergebnisse einer Regressionsschätzung verfälschen und müssen dementsprechend behandelt werden?

Zunächst einmal ist darauf zu verweisen, dass fehlende Werte nur dann überhaupt relevant werden, wenn sie die Variablen des zu schätzenden Regressionsmodells betreffen. Wird z.B. die Frage nach dem monatlichen Nettoeinkommen von vielen Befragten nicht beantwortet und ist die diesbezügliche Variable weder als Y- noch als X-Variable im zu schätzenden Regressionsmodell vertreten, dann sind die Fälle mit fehlenden Werten für die Regressionsanalyse irrelevant.

Genauso sind fehlende Werte nicht weiter behandlungsbedürftig, wenn sie als Folge einer beabsichtigten Filterführung bei der Befragung/Beobachtung entstanden sind. So weisen z.B. Variablen zu personalen Beschäftigungsverhältnissen (etwa dazu, ob eine ganz-, halbtägige oder geringfügige Beschäftigung vorliegt) naturgemäß eine große Anzahl von fehlenden Werten auf, weil oftmals in einem Survey auch Studierende, Arbeitslose und Hausfrauen/-männer befragt werden, die in keinem Beschäftigungsverhältnis stehen. Auch solche fehlenden Werte sind nicht behandlungsbedürftig, denn die dazugehörigen Fragen treffen für bestimmte Personen aus logischen Gründen ganz einfach nicht zu.

Mithin beschäftigen wir uns im Folgenden nur mit solchen fehlenden Werten, welche a) bei den Modellvariablen einer bestimmten Regressionsschätzung auftreten, und welche b) Modellvariablen betreffen, die bei bestimmten Untersuchungsfällen (Personen, Betriebe o.ä.) einen empirisch „wahren“ Wert aufweisen können, der aber (aus welchen Gründen auch immer) nicht erhoben werden konnte und somit im auszuwertenden Datensatz als „fehlend“ verbucht wurde. Dies bedeutet auch, dass Variablenwerte, mit denen Antwortkategorien wie z.B. „weiß nicht“ kodiert werden, nicht zwangsläufig als fehlende Werte betrachtet werden dürfen.

Zum Umgang mit regressionsanalytisch relevanten, fehlenden Werten stellt SPSS eine Vielzahl von Möglichkeiten zur Verfügung. Zu den diesbezüglichen Verfahren gehören unter anderem:

- (1) der listenweise Ausschluss von Fällen („listwise deletion“),
- (2) der paarweise Ausschluss von fehlenden Werten („pairwise deletion“),
- (3) die Imputation von Stichprobenmittelwerten,
- (4) die stochastische Regressionsimputation,
- (5) die Imputation nach EM-Schätzung („expectation maximization“),
- (6) die Multiple Imputation (MI-Verfahren).

Wir werden im Folgenden nicht alle diese sechs Strategien zur Behandlung von fehlenden Werten vorstellen.<sup>3</sup> Stattdessen werden wir uns auf drei Verfahren konzentrieren, die nach unserer Einschätzung die meisten Vorteile und die geringsten Risiken für regressionsanalytische Untersuchungen aufweisen. Diese ausgewählten Verfahren sind:

- der listenweise Ausschluss von Fällen („listwise deletion“),
- die stochastische Regressionsimputation,
- die Multiple Imputation (MI-Verfahren).

Diese Verfahren werden insbesondere deshalb ausgewählt, weil sich in der Regressionsanalyse die Problematik fehlender Werte und deren adäquate Behandlung ganz anders darstellen als in anderen statistischen Analysemodellen. Was das inhaltlich bedeutet, werden wir im SISS-Kapitel 3.1 „Der listenweise Ausschluss von Fällen“ ausführlich erläutern.

SPSS bietet drei der o.g. Strategien (1: „Listenweiser Fallausschluss“; 2: „Paarweiser Fallausschluss“ und 3: „Imputation von Stichprobenmittelwerten“) direkt innerhalb des Regressionsbefehls bzw. in der Menüsteuerung zur Durchführung von Regressionsanalysen als Verfahren zur Missing-Behandlung an. Die anderen Verfahren müssen außerhalb des Regressionsbefehls ausgeführt werden, und die Regressionsanalyse erfolgt dann erst in einem zweiten Schritt. Als Default-Einstellung verwendet SPSS bei der Regressionsanalyse die Strategie 1 („Listenweiser Fallausschluss“).

Die Strategie 2 („Paarweiser Fallausschluss“) und die Strategie 3 („Imputation von Stichprobenmittelwerten“) werden im Weiteren aus folgenden Gründen nicht weiter verfolgt:

(ad Strategie 2) Im Vergleich zum listenweisen Fallausschluss, stehen nach einem *paarweisen Fallausschluss* mehr Fälle zur Verfügung, da separat für jedes Variablenpaar so viele gültige Fälle wie möglich zur Koeffizientenschätzung herangezogen werden. Das Problem dabei ist jedoch, dass dadurch die Analyse-Fallzahl bei den verschiedenen Variablenpaaren extrem variieren kann. Kommen dann Fälle mit extremen Werten aufgrund von fehlenden Werten bei einzelnen Variablen nur bei manchen Koeffizientenschätzungen zum Einsatz, dann können die Ergebnisse dadurch stark verzerrt sein. Und auch die Schätzung von Standardfehlern und Konfidenzintervallen ist problematisch, wenn die Fallzahl innerhalb der Regressionsschätzung variiert (Schafer/Graham 2002: 155).

(ad Strategie 3) Wenn die fehlenden Werte einzelner Variablen durch deren *Mittelwerte* ersetzt werden, hat das automatisch zur Folge, dass die Varianz dieser Variablen reduziert wird. Denn allen Fällen mit fehlenden Werten werden pro Variable nur ein einziger, ganz bestimmter Wert (d.i. der Stichprobenmittelwert) zugeordnet. Und wenn es einen großen Anteil von Fällen mit fehlenden Wer-

---

<sup>3</sup> Neben diesen Strategien zum Umgang mit fehlenden Werten haben sich in der Forschungspraxis auch Schätzverfahren etabliert, die auf Maximum-Likelihood-(ML-)Algorithmen basieren. Ein Spezialfall davon ist die Imputation nach EM-Schätzung („expectation maximization“), die auch in SPSS zur Verfügung steht. In der Forschungspraxis hat sich jedoch bei der Behandlung von fehlenden Werten mit ML-Algorithmen weitgehend das Full Information Maximum Likelihood (FIML)-Schätzverfahren durchgesetzt, bei dem alle Informationen eines Datensatzes zur Schätzung der Modellparameter benutzt werden, ohne dass Fälle ausgeschlossen oder Werte imputiert werden müssten (vgl. Arbuckle 1996). Dieses Verfahren wird jedoch in SPSS nicht angeboten. Daher klammern wir die Strategien der ML-Schätzverfahren zur Behandlung von fehlenden Werten aus unserem Text aus.

ten bei einer bestimmten Variablen gibt, kann durch Mittelwert-Imputation eine ganz erhebliche Varianzreduktion ausgelöst werden. Bei einer Reduktion der Varianzen von Variablen werden aber auch deren Kovarianzen gemindert, was in aller Regel die Stärke der geschätzten Regressionskoeffizienten reduziert. Zudem verkleinern sich dabei oftmals auch die Standardfehler, was zu falschen Ergebnissen von Signifikanztests führen muss.<sup>4</sup> Folglich kann vor dem Gebrauch von Mittelwert-Imputationen bei fehlenden Werten nur gewarnt werden (insbesondere dann, wenn es eine große Anzahl von Fällen mit fehlenden Werten gibt). Die dabei auftretenden Risiken für verzerrte Schätzwerte sind so groß, dass auch dann, wenn die Ausfälle nicht rein zufällig entstanden sind, es oftmals vorzuziehen ist, alle Fälle mit fehlenden Werten aus der Regressionsanalyse auszuschließen anstatt Mittelwert-Imputationen durchzuführen.<sup>5</sup> Dafür sprechen auch die Argumente, die im folgenden SISS-Kapitel 3.1 vorgetragen werden.

Grundsätzlich sollte jeder Sozialforscher beim Thema „fehlende Werte“ das Folgende bedenken: Zwar sind die Techniken, die hier zum Umgang mit fehlenden Werten vorgestellt werden, sehr leistungsfähig und auch recht robust gegenüber möglichen Störeinflüssen. Aber:

- Der beste Umgang mit fehlenden Werten besteht immer noch darin, schon bei der Konzeptionierung und Durchführung einer Befragung/Beobachtung die Wahrscheinlichkeit für die Entstehung fehlender Werte möglichst gering zu halten. Z.B. kann häufig durch eine andere Fragetechnik (etwa indem die Einkommensfrage auch als Listenabfrage gestellt wird) die Anzahl fehlender Werte deutlich reduziert werden.
- Wenn in der statistischen Analyse zwischen mehreren Indikatoren gewählt werden kann (z.B. zwischen fünf Indikatoren zur Bildung eines Drei-Indikatoren-Index für die Messung des Sozioökonomischen Status) sollte auch darauf geachtet werden, dass nicht unbedingt derjenige Indikator mit der größten Anzahl fehlender Werte in die Modellkonstruktion einbezogen werden muss. Durch eine adäquate Variablenauswahl können manche Missing-Probleme bereits im Vorfeld der Regressionsanalyse vermieden werden.

## 2 Statistische Analyse fehlender Werte

Bevor in einer statistischen Untersuchung über den adäquaten Umgang mit fehlenden Variablenwerten nachgedacht wird, sollte zunächst die Quote von Fällen mit fehlenden Werten (Missingquote) ermittelt werden. Denn in der Forschungspraxis hat es sich bewährt, bei weniger als 5% von Fällen mit fehlenden Werten nur die kompletten Fälle für eine Regressionsschätzung zu benutzen und die anderen Fälle ohne weitere Behandlung aus der Analyse auszuschließen (das macht auch SPSS über die automatische Voreinstellung „listwise deletion“ von ganz alleine). Die Ergebnisverzerrung, die

---

<sup>4</sup> Bei unterschätzten Standardfehlern werden im Signifikanztest die p-Werte zu klein und die Konfidenzintervalle zu eng.

<sup>5</sup> Vgl. dazu Pigott 2001: 365; Tabachnick/Fidell 2013: 67.

von weniger als 5% fehlender Fälle ausgehen kann, ist als zu vernachlässigende Größe anzusehen (insbesondere angesichts der vielfältigen Risiken aller Imputationsverfahren).

Zudem kann in der Regressionsanalyse unter bestimmten, gar nicht einmal so selten anzutreffenden Bedingungen auch noch bei Missingquoten von weit über 5% ein listenweiser Ausschluss der Fälle mit fehlenden Werten die beste Strategie sein. Dies ist eine Besonderheit der Regressionsanalyse beim Umgang mit fehlenden Werten, die sie von anderen statistischen Analyseverfahren unterscheidet (mehr dazu im folgenden SISS-Kapitel 3.1).

Zu beachten ist, dass sich die 5%-Daumenregel sowohl auf die fehlenden Werte bei jeder einzelnen Modellvariablen als auch auf die kumulierte Missingquote bezieht. Denn wenn z.B. ein Regressionsmodell nur drei Variablen umfasst ( $Y$ ,  $X_1$  und  $X_2$ ) und bei jeder Variablen „nur“ 4% aller Fälle auch fehlende Werte aufweisen, kann daraus schnell eine kumulierte Missingquote von über 10% entstehen (maximal sogar von 12%). Hohe kumulierte Missingquoten entstehen immer dann, wenn bei jeder Variablen eine andere Gruppe von befragten Personen aufgrund von Antwortverweigerungen die fehlenden Werte erzeugt. Dies wird in der folgenden Tabelle 1 verdeutlicht. Dort fehlen bei jeder der drei Modellvariablen die Werte von drei Personen. Das ergibt pro Variable eine Missingquote von nur 3%. Da es aber bis auf eine einzige Person (Person Nr. 9) stets unterschiedliche Befragte sind, die bestimmte Fragen nicht beantwortet haben, beträgt die Missingquote für den gesamten Variablensatz der Regressionsschätzung bereits 7%, denn 7 von insgesamt 10 Personen weisen zumindest eine Variable mit einem fehlenden Wert auf (ausgenommen sind nur die Personen Nr. 7, 8 und 10).

**Tabelle 1:** Ausgeschlossene Fälle (Personen) und fehlende Werte bei drei Modellvariablen (fiktive Zahlen)

Person Nr.	Y	$X_1$	$X_2$
①	○	3	7
②	○	2	6
③	17	○	3
④	20	○	2
⑤	24	5	○
⑥	25	6	○
7	14	1	2
8	16	1	1
⑨	○	○	○
10	29	2	2

Mit der folgenden SPSS-Syntax 1 lässt sich leicht die Missingquote „missk“ für die Gesamtheit aller Variablen eines Regressionsmodells („k“ symbolisiert die Gesamtheit aller Modellvariablen) ermitteln:<sup>6</sup>

#### SPSS-Syntax 1

```
COUNT missk = y x1 x2 (MISSING).  
IF (missk>0) missk=1.  
FREQ missk.
```

Zur Veranschaulichung dieser SPSS-Syntax und auch zur Veranschaulichung der folgenden statistischen Analysen wollen wir ein Beispiel aus der Antisemitismusforschung benutzen. Uns interessiert dabei der Effekt von Bildung auf das Ausmaß von Antisemitismus bei Kontrolle des Alterseffekts.

Zur Messung von Antisemitismus haben wir einen additiven Index (ungewichtet, gemittelt) aus den Bewertungen von vier Items gebildet: a) „Juden haben auf der Welt zu viel Einfluss“, b) „Ich empfinde Scham über deutsche Untaten an Juden“ (gedreht), c) „Juden nutzen die deutsche Vergangenheit aus“, d) „Juden sind an Verfolgungen nicht unschuldig“. Zur Konstruktion des Index „AS\_Index“ wurde die folgende SPSS-Syntax 2 benutzt:

#### SPSS-Syntax 2

```
COMPUTE AS_Index = MEAN.4(item1, item2, item3, item4).
```

Die Prädiktorvariable „Bildung (allgemeine Schulbildung)“ wird in diesem Beispiel metrisch definiert (vgl. dazu den nächsten Textabsatz). Die Variable erhielt die Werte: „1=ohne Abschluss“, „2=Volks-, Hauptschulabschluss“, „3=Mittlere Reife o.ä.“, „4=Fachhochschulreife o.ä.“, „5= Hochschulreife“. Die Ausprägungen „noch Schüler“ und „anderer Abschluss“ wurden als fehlende Werte („missings“) definiert. Die Kontrollvariable „Alter“ misst die Lebensjahre der Befragten. Als empirische Datenbasis wurde eine 3%-Zufallsauswahl (N=92) von Fällen ohne Missings aus den Daten der allgemeinen Bevölkerungsumfrage „ALLBUS 2012“ (N=3480) gezogen. Auf diese Weise erhielten wir (allein zu Vergleichszwecken) ein komplettes Sample ohne fehlende Werte bei den drei Variablen „AS\_Index“, „Bildung“ und „Alter“ (Datensatz „AS0.sav“). Aus diesem Datensatz wurde dann der neue Datensatz „AS1.sav“ gebildet, indem bei insgesamt 15 durch Zufall ausgewählten Befragten (15 aus 92 = 16%) der „AS\_Index“-Wert gelöscht und durch „Missing“ ersetzt wurde.

Bei der metrischen Definition von Bildung gehen wir davon aus, dass bei dieser Variablen vier wichtige Voraussetzungen zur metrischen Skalierung von ordinal gemessenen Variablen erfüllt sind: (1) Die Messskala der Variablen umfasst möglichst viele und nicht weniger als fünf Skalenpunkte. (2) Die Abstände zwischen den benachbarten Skalenpunkten dieser Variablen sind numerisch und semantisch äquivalent. (3) Die Messwerte dieser Variablen „X“ haben einen Bezug zu einer kontinuierlich verteilten, latenten X\*-Hintergrundvariablen. (4) Die Verteilung der Variablenwerte muss zwar nicht

<sup>6</sup> Diese SPSS-Syntax ermittelt sowohl Fälle mit benutzerdefinierten als auch mit systembedingten fehlenden Werten (SYSMIS) in SPSS-Datensätzen.



normalverteilt sein, ist aber auch nicht extrem schief und weist eine nicht zu kleine Varianz auf. Allerdings sei hier darauf hingewiesen, dass die Gültigkeit der Voraussetzungen (2) und (3) für die Variable „Bildung“ in der sozialwissenschaftlichen Forschung umstritten ist und deshalb oftmals Bildung als nicht metrisierbar angenommen wird. Da es sich hier allerdings um ein rein didaktisch begründetes Anwendungsbeispiel handelt, wollen wir in unseren Ausführungen ohne weitere Problematisierung von einer metrischen Skalierbarkeit der Variablen „Bildung“ ausgehen.

Die folgende Tabelle 2 zeigt die kumulierte Missingquote über die Variablen „Antisemitismus-Index“, „Bildung“ und „Alter“ im von uns benutzten Datensatz „AS1.sav“ (=Antisemitismusdaten, s.o.). Sie beträgt 16,3% und wurde mit der SPSS-Syntax 3 ermittelt. Zur Berechnung des Index wurde die oben abgedruckte SPSS-Anweisung 2 benutzt. Allerdings wurden die fiktiven Variablennamen, die noch oben in der SPSS-Anweisung zur Konstruktion eines additiven Index benutzt worden sind, durch die Variablennamen aus dem Antisemitismusdatensatz ersetzt.

### SPSS-Syntax 3

```
COUNT missk= AS_Index, Bildung, Alter (MISSING).
IF (missk > 0) missk = 1.
FREQ missk.
```

Tabelle 2: Kumulierte Missingquote (Antisemitismusdaten)

	missk		Gültige Prozent	Kumulative Prozente
	Häufigkeit	Prozent		
Gültig ,00	77	83,7	83,7	83,7
1,00	15	16,3	16,3	100,0
Ges.summe	92	100,0	100,0	

Wie Tabelle 2 zeigt, beträgt die kumulierte Missingquote für den Datensatz mit den Modellvariablen „AS\_Index“, „Bildung“ und „Alter“ insgesamt 16,3%. Die separaten Missingquoten für jede einzelne Variable im Antisemitismusmodell lassen sich mit der folgenden SPSS-Syntax 4 abrufen:

### SPSS-Syntax 4

```
RECODE AS_Index Bildung Alter (MISSING=1) (ELSE=0) INTO
miss1 miss2 miss3.
EXECUTE.
FREQ VARIABLES=miss1 miss2 miss3.
```

Die drei mit SPSS-Syntax 4 erstellten SPSS-Tabellen informieren über die Missingquote für jede einzelne Variable im Antisemitismusmodell. In der folgenden Tabelle 3 werden diese Quoten noch einmal aufgelistet. Da in unserem Beispiel die Variablen „Bildung“ und „Alter“ keine fehlenden Werte aufweisen, ist die kumulierte Missingquote identisch mit der Missingquote bei der Variablen „AS\_Index“:

Tabelle 3: Variablenspezifische Missingquoten (Antisemitismusdaten)

Variable	Missingquote
AS_Index	16,3%
Bildung	0,0%
Alter	0,0%

Liegt die kumulierte Missingquote für die Variablen eines zu analysierenden Modells oberhalb von 5%, sollte die Systematik der Fälle mit fehlenden Werten statistisch untersucht werden (s.u.). Denn von dieser Systematik kann es abhängen, welche der drei im Folgenden vorgestellten Strategien zum Umgang mit fehlenden Werten (vgl. SISS-Kapitel 3.1 bis 3.3) in einer Regressionsanalyse eingesetzt werden sollte.

Zur Bestimmung der Systematik fehlender Werte werden in der Forschungspraxis drei verschiedene „Ausfallmechanismen“ (für die jeweils unterschiedliche Annahmen gelten) unterschieden. Diese werden bezeichnet als:

- (1) MCAR („missing completely at random“)
- (2) MAR („missing at random“)
- (3) MNAR („missing not at random“)<sup>7</sup>

(ad MCAR) Sind die fehlenden Werte das Ergebnis eines MCAR-Ausfallmechanismus, so sind sie – so die Annahme – absolut zufällig entstanden. Demnach dürfen fehlende Werte bei einer Variablen weder aufgrund von Gründen, die in anderen Modellvariablen liegen, noch aufgrund von Eigenschaften der Variablen mit fehlenden Werten selbst entstanden sein. Letzteres bedeutet, dass die „wahren“ bzw. verschwiegenen bzw. nicht zu ermittelnden Variablenwerte gleichmäßig über alle Ausprägungen der Variablen verteilt sein müssten. Ist dies nicht der Fall, dann würden die Missings gehäuft in bestimmten Ausprägungsbereichen der Variablen vorkommen und die MCAR-Annahme wäre verletzt. Dies wäre z.B. dann gegeben, wenn alle Mitglieder einer radikalen politischen Partei diese Mitgliedschaft verheimlichen würden und bei einer Befragung zu Parteien-Mitgliedschaften ihre Antwort

<sup>7</sup> Alternativ zur Bezeichnung „MNAR“ findet man in der Literatur auch die Bezeichnungen „NMAR“ (not missing at random) und „NI“ (non ignorable).

verweigern würden. Ein anderes Beispiel wäre, wenn bei Einkommensfragen besonders arme und besonders reiche Personen ihre Einkommensangaben verweigerten. In gleicher Weise dürfen bei einem MCAR-Ausfallmechanismus auch die Ausprägungen der anderen beobachteten Variablen keinen Einfluss darauf haben, ob bestimmte Variablenwerte fehlen oder nicht fehlen, denn die fehlenden Werte einer Variablen müssen rein zufällig entstanden sein. Im Einkommensbeispiel würde dies bedeuten, dass ein fehlender Wert bei der Einkommensfrage z.B. nicht von der Schulbildung und vom Alter der Person abhängen darf. Dementsprechend sind die Annahmen zum MCAR-Ausfallmechanismus sehr restriktiv und dürften empirisch nicht allzu häufig zu beobachten sein.

(ad MAR) Auch beim MAR-Ausfallmechanismus sind die fehlenden Werte einer Variablen ebenso wie beim MCAR-Mechanismus nicht von den Ausprägungen der tatsächlichen (aber „verheimlichten“) Werte der betroffenen Variablen abhängig. Treten beispielsweise fehlende Werte bei der Einkommensvariablen auf, so wird angenommen, dass die fehlenden Werte nicht dadurch zustande kommen, dass Personen, die ein sehr hohes (oder auch sehr niedriges) Einkommen beziehen, dieses eher verschweigen als Personen mit einem durchschnittlichen Einkommen. Allerdings haben beim MAR-Mechanismus andere beobachtete Variablen einen direkten Einfluss auf die Ausfallwahrscheinlichkeit. So kann z.B. die verweigerte Antwort auf die Frage nach der Einkommenshöhe vom Alter des Befragten beeinflusst sein, wenn dieser aufgrund seiner Zugehörigkeit zu einer älteren Kohorte noch eine rigide Vorstellung von Privatsphäre hat als jüngere Befragte. Insofern bedeutet MAR zwar auch „zufällig entstanden“, aber nur zufällig hinsichtlich der von fehlenden Werten betroffenen Variablen, wobei diese Zufälligkeit auch dann bestehen bleibt, wenn die Effekte anderer beobachteter Variablen auf die Wahrscheinlichkeit fehlender Werte kontrolliert werden.

(ad MNAR): Beim MNAR-Ausfallmechanismus wird angenommen, dass ein fehlender Variablenwert a) wie auch schon beim MAR-Ausfallmechanismus von den Effekten anderer Variablen im Datensatz abhängig sein kann, und b) dass der fehlende Wert aber auch von der tatsächlichen (aber unbekannten bzw. fehlenden) Ausprägung der betroffenen Variablen selbst verursacht sein kann. Diese beiden Annahmen sind somit weniger restriktiv als die Annahmen zum MCAR- oder MAR-Mechanismus. Im o.g. Einkommensbeispiel bedeutet dies, dass nach den Annahmen des MNAR-Ausfallmechanismus die Wahrscheinlichkeit fehlender Einkommensangaben sowohl von der Höhe des Einkommens (z.B. Ausfälle bei besonders hohem Einkommen) als auch von den Einflüssen anderer beobachteter Variablen (z.B. von Geschlecht, Alter usw.) abhängig sein kann.

Das folgende Beispiel und die Tabelle 4 machen die Unterschiede zwischen MCAR-, MAR-, MNAR-Ausfallmechanismen deutlich:<sup>8</sup>

Im Beispiel wird angenommen, dass von 20 Beschäftigten eines Unternehmens die Beschäftigungsdauer (in Arbeitstagen) bekannt sei. Zudem wird angenommen, dass die Bewertung der Arbeitsleistung eines jeden Beschäftigten unter vier verschiedenen Bedingungen vorläge. Diese Bedingungen sind (vgl. dazu auch Tabelle 4): a) es gibt keine fehlenden Werte bei der Angabe der Arbeitsleistung (Spalte „keine Missings“), b) es gibt einen MCAR-Ausfallmechanismus bei der Arbeitsleistung (Spalte „MCAR“), c) es gibt einen MAR-Ausfallmechanismus (Spalte „MAR“), d) es gibt einen MNAR-Ausfallmechanismus (Spalte „MNAR“). Die mittlere Beschäftigungsdauer aller 20 Personen beträgt 100,00 Arbeitstage. Mit diesen Daten können separat für jede Spalte b), c) und d) die mittleren Beschäftigungszeiten für die Personengruppe ohne Missings und für die Personengruppe, bei der die Angabe der Arbeitsleistung fehlt (Missings), ermittelt und miteinander verglichen werden.

Unter der MCAR-Bedingung fehlen die Messwerte zur Arbeitsleistung rein zufällig. Deshalb wird zur Generierung der fehlenden Werte in der Spalte „MCAR“ ein Zufallsmechanismus benutzt. Die fehlenden Werte werden weder durch die Bewertung der Arbeitsleistung noch durch die Beschäftigungsdauer beeinflusst. Fehlende Werte gibt es bei jedem Ausmaß von Arbeitsleistung (vgl. Spalte „keine Missings“) und fehlende Werte gibt es bei langer, mittlerer und kurzer Beschäftigung. So könnten hier die fehlenden Werte z.B. dadurch entstanden sein, dass Beschäftigte aufgrund einer Erkrankung nicht an den entsprechenden Leistungstests teilgenommen haben. Weil der Ausfall von Werten rein zufällig entsteht, unterscheiden sich in der Spalte „MCAR“ die Mittelwerte der Arbeitstage für die beiden Gruppen nicht wesentlich voneinander. Sie betragen 99,73 Tage (ohne Missings) und 100,80 Tage (nur Missings).

Anders sieht es beim Vergleich der mittleren Beschäftigungszeiten in der Spalte „MAR“ aus. Dort betragen die Mittelwerte der Arbeitsdauer 105,47 Tage (nur gültige Werte) und 83,69 Tage (nur Missings). Die beiden Personengruppen ohne und mit fehlenden Werten unterscheiden sich unter MAR-Bedingungen also sehr deutlich voneinander. Dies kann auch nicht überraschen, wenn man die Spalte „MAR“ etwas näher betrachtet. Fehlende Leistungswerte gibt es nur bei denjenigen Beschäftigten, die erst seit Kurzem im Unternehmen beschäftigt sind.

Die letzte Spalte von Tabelle 4 (Spalte „MNAR“) enthält fehlende Werte, deren Auftreten sowohl von der Beschäftigungsdauer als auch von der Bewertung der Arbeitsleistung abhängt. Fehlende Werte kommen dabei besonders häufig bei mittellangen Beschäftigungen und bei geringen Leistungswerten vor (vgl. die entsprechenden Werte in der Spalte „keine Missings“). Vergleicht man wiederum die Mittelwerte der Beschäftigungstage in der Gruppe mit vorhandenen Leistungswerten und in der Gruppe mit fehlenden Leistungswerten, so zeigt sich auch dort der erwartete Unterschied: Die entsprechenden Mittelwerte betragen 102,8 Tage (nur gültige Werte) und 91,6 Tage (nur Missings).

---

<sup>8</sup> Das Beispiel und die Tabelle 4 wurden nach Enders 2010: 6-8 entwickelt.

Tabelle 4: Konsequenzen unterschiedlicher Ausfallmechanismen (fiktives Beispiel)

Person	Arbeits- tage	Bewertung der Arbeitsleistung			
		a) keine Missings	b) MCAR	c) MAR	d) MNAR
1	78	9	-	-	9
2	84	13	13	-	13
3	84	10	-	-	10
4	85	8	8	-	-
5	87	7	7	-	-
6	91	7	7	7	-
7	92	9	9	9	9
8	94	9	9	9	9
9	94	11	11	11	11
10	96	7	-	7	-
11	99	7	7	7	-
12	105	10	10	10	10
13	105	11	11	11	11
14	106	15	15	15	15
15	108	10	10	10	10
16	112	10	-	10	10
17	113	12	12	12	12
18	115	14	14	14	14
19	118	16	16	16	16
20	134	12	-	12	12
mittlere Beschäft.- dauer (Tage)					
- ohne Miss.:	100,00		99,73	105,47	102,80
- nur Miss.:	(---)		100,80	83,69	91,60

Von den drei hier vorgestellten Ausfallmechanismen lässt sich allein der MCAR-Mechanismus relativ eindeutig statistisch überprüfen. Die diesbezüglichen Tests können jedoch auch bei diesem Mechanismus keine Gewissheit über die Gültigkeit aller seiner Annahmen vermitteln. Denn es lässt sich immer nur etwas testen, was auch beobachtet werden kann. Dies ist aber gerade bei fehlenden Werten nicht möglich. Deshalb folgen die Tests der MCAR-Eigenschaft einer Logik, die auf die Falsifikation einer MCAR-Annahme ausgerichtet ist. Somit kann mit diesen Tests zwar die Gültigkeit der MCAR-Eigenschaft widerlegt aber nicht bestätigt werden.

Getestet wird der MCAR-Ausfallmechanismus häufig mit dem sog. „Little-Test“. Dieser Test informiert darüber, ob der Zusammenhang zwischen den fehlenden und den gemessenen Werten aller Modellvariablen als rein zufällig zu betrachten ist.

Beim multivariaten Little-Test werden die Mittelwerte für jede Modellvariable in der Untersuchungsgruppe (Gruppe mit fehlenden Werten) mit den Mittelwerten in der Kontrollgruppe (Gruppe ohne fehlende Werte) verglichen. Der Test ist somit äquivalent zu multiplen t-Tests mit einer bivariaten

Variablen.<sup>9</sup> Getestet wird die Nullhypothese, dass die MCAR-Eigenschaft erfüllt ist. D.h., dass bei einem signifikanten Ergebnis des Little-Tests diese Nullhypothese zu verwerfen wäre, denn dann wäre die Irrtumswahrscheinlichkeit für eine solche Entscheidung sehr klein (unter 5%). Also sollte, um diese Nullhypothese beizubehalten, die Irrtumswahrscheinlichkeit „p“ möglichst groß sein und am besten über 10% liegen (auf jeden Fall aber gleich oder größer als 5% bzw. 0,05 sein). Im Little-Test wird also ein nicht-signifikantes Ergebnis gesucht.

In SPSS wird der Little-Test für unsere Modellvariablen mit folgender Syntax aufgerufen:

#### SPSS-Syntax 5

MVA VARIABLES = AS\_Index, Bildung, Alter /EM.

Mit der folgenden Tabelle 5 wird in SPSS über die Ergebnisse des Little-Tests informiert (zu finden unterhalb der diversen EM-Tabellen in der SPSS-Ausgabe).<sup>10</sup>

Tabelle 5: Ergebnisse des Little-Tests (Antisemitismusdaten)

MCAR-Test nach  
Little: Chi-Quadrat =  
3,020, DF = 2, Sig.  
= ,221

Die beim Little-Test ausgegebene Chi-Quadrat-Statistik (vgl. Tabelle 5) lässt sich also nutzen, um zu testen, ob im Datensatz bestimmte Werte in rein zufälliger Weise und unabhängig von den anderen Modellvariablen fehlen. Nach dem Ergebnis des hier abgedruckten Tests ist die MCAR-Eigenschaft hinsichtlich unserer Daten zu akzeptieren, da die diesbezügliche Irrtumswahrscheinlichkeit mit einem Wert von 0,221 ausreichend groß ist und deshalb nur mit einem hohen (bzw. zu hohen) Risiko die Nullhypothese (= Gültigkeit der MCAR-Annahmen) verworfen werden könnte.

---

<sup>9</sup> Zum Little-Test vgl. Little 1988 (insbesondere S. 1198). Ausf. dazu auch: Enders 2010: 19-21, 32-35. Zur Logik von t-Tests vgl. Field 2013: 364ff.

<sup>10</sup> Auf die Bedeutung der EM-Tabellen in der SPSS-Ausgabe wird im vorliegenden SISS-Beitrag nicht eingegangen. Vgl. dazu Cohen et al. 2003: 440-442; Enders 2010: 86-163.

Folgendes ist bei Durchführung des Little-Tests zu beachten:

- a) Der Test ist asymptotisch, eignet sich also am besten für Datensätze mit großen Fallzahlen. In der Forschungspraxis wird er jedoch häufig auch bei kleineren Samples eingesetzt.<sup>11</sup>
- b) Der Test eignet sich nur für metrische Variablen (was in unserem Antisemitismusbeispiel bei den Variablen „Antisemitismus-Index“, „Bildung“ und „Alter“ vorausgesetzt wird, was allerdings bei der Bildungsvariablen auch strittig sein kann, da dort die Abstände zwischen den Bildungsabschlüssen als gleichgroß interpretiert werden müssen, um ein metrisches Skalenniveau annehmen zu können). Für kategoriale Variablen, wie z.B. die nominale Variable „Geschlecht“, ist der Little-Test nicht geeignet.
- c) Die metrischen Variablen sollten im Idealfalle zumindest annäherungsweise normalverteilt sein. Allerdings ist der Test recht robust gegenüber Verteilungen, die von der Normalverteilungsform abweichen.<sup>12</sup>
- d) Der Test kann auch bei hohen kumulierten Missingquoten (z.B. 25%) und mit einer großen Anzahl von Variablen durchgeführt werden.

Das Ergebnis des Little-Tests kann auch noch mit weiteren Tests abgesichert werden. Dazu empfehlen wir allerdings nicht die von SPSS angebotenen bivariaten t-Tests, denn deren Logik ist nicht grundsätzlich verschieden von der Logik des Little-Tests.

Stattdessen schlagen wir dazu eine logistische Regressionsschätzung mit einer binär kodierten Missing-Variablen als abhängige Variable vor.<sup>13</sup> Damit lässt sich überprüfen, ob zumindest ein Teil der MCAR-Annahmen, nämlich die Unabhängigkeit der Wahrscheinlichkeit fehlender Werte von Einflüssen anderer Modellvariablen, erfüllt ist oder nicht. Denn wenn in einer logistischen Regression mit der Missing-Variablen als binärer abhängiger Variablen zumindest ein signifikanter Prädiktoreffekt nachgewiesen wird, so ist damit die MCAR-Eigenschaft fehlender Werte widerlegt.

Für eine logistische Regression zur zusätzlichen Überprüfung der MCAR-Eigenschaft sprechen einige gute Gründe:

- a) Es lassen sich damit auch die Effekte von kategorialen Modellvariablen auf die Wahrscheinlichkeit fehlender Werte überprüfen.

---

<sup>11</sup> Vgl. Li 2013: 804.

<sup>12</sup> Vgl. Li 2013: 797.

<sup>13</sup> Zur Durchführung logistischer Regressionsschätzungen vgl. Urban 1993; Urban/Mayerl 2011.

- b) Die simultane Schätzung von Effekten im multivariaten Modell kann auch solche Einflüsse auf die Wahrscheinlichkeit fehlender Werte aufspüren, die in bivariaten Modellen von den Effekten anderer Modellvariablen überlagert sind und deshalb dort nur schwerlich entdeckt werden können.
- c) Die eher problematischen Verteilungsannahmen von t-Tests (z.B. Varianzgleichheit), auf denen auch der Little-Test beruht, gelten in der logistischen Regressionsschätzung nicht.
- d) Eine logistische Regressionsanalyse kann zumindest einen Teil der MAR-Annahmen testen. Denn wenn sie Effekte von unabhängigen Variablen auf die Wahrscheinlichkeit von fehlenden Werten nachweist, sind auf jeden Fall die MCAR-Annahmen nicht erfüllt und es können als Ausschlussmechanismen nur noch MAR oder NMAR in Frage kommen.

Zur Durchführung einer logistischen Regressionsanalyse für eine Überprüfung der MCAR-Eigenschaft bzw. für eine Absicherung von Ergebnissen des Little-Tests müssen zunächst eine oder mehrere abhängige binäre Missing-Variable(n) gebildet werden. Dazu werden diejenigen Modellvariablen, die eine hohe oder eine zu hohe Missingquote aufweisen, rekodiert. Wenn ein Untersuchungsfall (Person, Organisation o.ä.) bei einer dieser Variablen einen fehlenden Wert aufweist, wird ihr der Wert „1“ zugewiesen, ansonsten erhält sie den Wert „0“. Dies lässt sich in unserem Datensatz für die Variable „AS\_Index“ mit den folgenden SPSS-Syntax 6 durchführen. Dabei wird der neu gebildeten Missing-Variablen der Name „AS\_miss“ zugewiesen:

#### SPSS-Syntax 6

```
RECODE AS_Index (MISSING=1) (ELSE=0) INTO AS_miss.  
EXECUTE.  
FREQ VARIABLES=AS_miss.
```

Im Anschluss an die Rekodierung der Variablen mit fehlenden Werten sollte die logistische Regressionsschätzung durchgeführt werden. Dabei wird die binäre Variable „AS\_miss“ als abhängige Variable benutzt. Als Prädiktoren (unabhängige Variablen) sollten alle Modellvariablen mit Ausnahme der Variablen, deren Wert zur Definition der Missing-Variablen benutzt wurden (hier: „AS\_Index“) in die Schätzung einbezogen werden<sup>14</sup>. In unserem Antisemitismusmodell sind das die unabhängigen Variablen „Bildung“ und „Alter“. Die dementsprechende Regressionsschätzung wird für unser Beispiel mit der folgenden SPSS-Syntax 7 aufgerufen:

---

<sup>14</sup> Allerdings sollten Modellvariablen, die selbst eine hohe bis sehr hohe Missingquote aufweisen, nicht als Prädiktoren in dieses Modell einbezogen werden.



SPSS-Syntax 7

```
LOGISTIC REGRESSION VARIABLES AS_miss
/METHOD=ENTER Bildung Alter.
```

Die folgende Tabelle 6 zeigt die Ergebnisse der logistischen Regressionsschätzung zur Absicherung der Resultate des Little-Tests bzw. zum Test der MCAR-Eigenschaft. Danach hat weder der Prädiktor „Bildung“ noch der Prädiktor „Alter“ einen signifikanten Effekt auf die (logarithmierte) Chance für die Existenz fehlender Werte bei der Variablen „AS\_Index“ (bzw. bei deren rekodierter Schwestervariablen „AS\_miss“). Somit kann das Ergebnis des hier geschätzten logistischen Regressionsmodells das Resultat des Little-Tests bestätigen. Die Wahrscheinlichkeit fehlender Werte beim Antisemitismus-Index ist unabhängig von Einflüssen der anderen Prädiktorvariablen. Daraus kann gefolgert werden, dass die fehlenden Werte von „AS\_Index“ durchaus die MCAR-Eigenschaft aufweisen können.

Tabelle 6: Logistische Regressionsschätzung (Antisemitismusdaten)

Variablen in der Gleichung						
	B	Standard- fehler	Wald	df	Sig.	Exp(B)
Schritt 1 <sup>a</sup> Bildung	,370	,240	2,376	1	,123	1,447
Alter	-,003	,018	,030	1	,862	,997
Konst.	-2,702	1,401	3,720	1	,054	,067

Wenn jedoch (anders als in unserem Beispiel) die MCAR-Eigenschaft von fehlenden Werten durch Little-Test und/oder durch logistische Regressionsschätzung widerlegt wurde, ist analytisch zu überlegen, ob die fehlenden Werte das Ergebnis eines MAR- oder eines MNAR-Ausfallmechanismus sein könnten. Zwischen beiden Möglichkeiten kann allerdings nicht mittels eines statistischen Tests entschieden werden. Denn beide gehen von einem mittels Prädiktoren „fremdgesteuerten“ Fehlen bestimmter Variablenwerte aus. Beide Ausfallmechanismen unterscheiden sich mithin nur dadurch voneinander, dass der MAR-Mechanismus eine Zufälligkeit fehlender Werte bei Kontrolle von Prädiktoreffekten annimmt, während der MNAR-Mechanismus auch diese Zufälligkeit ausschließt. Da darüber aber nicht empirisch zu entscheiden ist (weil die „wahren“ Ausprägungen von fehlenden Werten unbekannt sind) hilft bei diesem Problem auch kein statistischer Test weiter.

Allerdings lässt sich, wie oben gezeigt, immerhin eine notwendige (jedoch auch nicht hinreichende, s.o.) Voraussetzung für beide Ausfallmechanismen statistisch testen: das Nichtvorhandensein von MCAR. Deshalb schlagen wir vor, bei jeder statistischen Analyse fehlender Werte auch die oben

erläuterte logistische Regressionsschätzung einzusetzen. Denn diese Schätzung kann drei verschiedene Funktionen übernehmen:<sup>15</sup>

- a) Sie kann die MCAR-Eigenschaft statistisch widerlegen, bzw. (im Umkehrschluss) bei nicht nachzuweisenden Prädiktoreffekten eine Akzeptanz der MCAR-Annahmen unterstützen.
- b) Sie kann das Ergebnis des Little-Test zusätzlich absichern.
- c) Sie kann die Existenz einer notwendigen (aber nicht hinreichenden) Voraussetzung für die Gültigkeit des MAR- oder MNAR-Ausfallmechanismus überprüfen.

### **3 Strategien zum Umgang mit fehlenden Werten in der Regressionsanalyse**

Nachdem die statistische Analyse der fehlenden Werte unter den Variablen eines Regressionsmodells durchgeführt worden ist (vgl. SISS-Kapitel 2), kann damit begonnen werden, entweder die fehlenden Werte aus der Analyse auszuschließen oder die fehlenden Werte durch statistisch geschätzte Werte zu ersetzen. Dazu stellt SPSS verschiedene Verfahren/Strategien zur Verfügung (vgl. SISS-Kapitel 1). Wie zuvor bereits erwähnt, möchten wir von diesen Verfahren insbesondere drei Strategien für die praktische Durchführung von Regressionsanalysen empfehlen. Diese drei Strategien sind:

- der listenweise Ausschluss von Fällen („listwise deletion“),
- die stochastische Regressionsimputation,
- die Multiple Imputation (MI-Verfahren).

Diese drei Verfahren werden in den folgenden SISS-Kapiteln 3.1 bis 3.3 im Einzelnen vorgestellt.

#### **3.1 Der listenweise Ausschluss von Fällen („listwise deletion“)**

Die wohl am häufigsten gewählte Strategie bei der Behandlung fehlender Werte ist der komplette Ausschluss einzelner Untersuchungsfälle (Personen, Organisationen o.ä.), wenn diese Fälle mindestens auf einer Modellvariablen einen fehlenden Wert aufweisen. In unserem Beispiel aus Tabelle 1 (SISS-Kapitel 2) wären das die sieben Personen No. 1 bis 6 und No. 9.

Ein solcher Ausschluss ganzer Fälle wird in SPSS als „listwise deletion“ bezeichnet.<sup>16</sup> SPSS benutzt das „listwise deletion“-Verfahren als Voreinstellung. Wenn SPSS keine anderen Anweisungen vom Anwender erhält, werden von SPSS nur komplette Fälle in einer Regressionsanalyse berücksichtigt.

---

<sup>15</sup> Natürlich bedingen sich diese Funktionen auch teilweise gegenseitig.

<sup>16</sup> Die Bezeichnung "listwise deletion" bedeutet, dass jeder fehlende Werte bei jeder einzelnen Variablen, die auf der Liste steht, d.h. die im Regressionsmodell berücksichtigt wird, zum Ausschluss des dazugehörigen, ganzen Untersuchungsfalls führt.

Der listenweise Ausschluss von Fällen setzt üblicherweise die MCAR-Eigenschaft der fehlenden Werte voraus. Wenn diese erfüllt ist (zur statistischen Überprüfung dieser Eigenschaft vgl. SISS-Kapitel 2), können die zum fehlenden Wert gehörigen Untersuchungsfälle relativ konsequenzenlos aus der Analyse ausgeschlossen werden. Denn durch den Ausschluss wird der zu analysierende Datensatz allein um Fälle mit zufällig entstandenen fehlenden Werten reduziert. Die ausgeschlossenen Fälle können dann als Zufallsauswahl aus dem Original-Datensatz betrachtet werden. Bei einem tatsächlichen MCAR-Ausschlussmechanismus werden die geschätzten Regressionsparameter und die benutzten Test-Statistiken nicht verfälscht, auch wenn die Missingquote hoch ist (z.B. 25 oder 40 Prozent).

Allerdings kann bei hohen Missingquoten ein listenweiser Ausschluss fehlender Fälle dann problematisch werden, wenn die Fallzahl des auszuwertenden Datensatzes klein ist (z.B. bei einem  $N < 200$ ). Denn bei einer durch Fallausschluss reduzierten Fallzahl reduziert sich auch die Teststärke der jeweiligen Datenanalyse, sodass es schwierig werden kann, bei Effekten mit geringer Stärke auch fehlerminimale Signifikanztests durchzuführen (mögliche Signifikanzen werden dann überhaupt nicht oder nur mit einer zu großen Irrtumswahrscheinlichkeit identifiziert).

Zudem wirkt ein Ausschluss vieler Fälle für Studien mit geringen Fallzahlen besonders tragisch, weil generell bei einem listenweisen Ausschluss aufgrund der Fallreduktion weniger empirische Information für die Regressionsschätzung zur Verfügung steht und deshalb der Standardfehler größer und damit die Entdeckung von Signifikanzen schwieriger wird.

Für die Forschungspraxis ist Folgendes besonders wichtig: Generell betrachtet (und einmal abgesehen vom oben erläuterten Fallzahlproblem) gibt es für Regressionsanalysen einen ganz besonderen Vorteil beim listenweisen Ausschluss fehlender Fälle. Denn erfreulicherweise kann in Regressionsanalysen bei bestimmten Voraussetzungen nicht nur unter MCAR-Bedingungen, sondern auch unter MAR-Bedingungen und auch unter noch weniger restriktiven Bedingungen ein listenweiser Ausschluss durchgeführt werden. Und dann können in Regressionsanalysen auch die erheblichen Unsicherheiten und Risiken, die mit den üblichen Imputationsverfahren verbunden sind (mehr dazu im Folgenden), vermieden werden.<sup>17</sup>

---

<sup>17</sup> Diese Einschätzung entspricht auch der Bewertung durch Paul D. Allison (ders. 2002: 5f).

Es stellt sich mithin die Frage: Unter welchen Voraussetzungen kann selbst bei Nicht-MCAR-Bedingungen ein listenweiser Ausschluss von Fällen mit fehlenden Werten in einer Regressionsanalyse durchgeführt werden?

In der Methodenliteratur werden dafür drei spezielle (weil nur für Regressionsanalysen gültige) Anwendungsvoraussetzungen genannt:<sup>18</sup>

### **(1) Die Y-bezogene Voraussetzung**

Wenn in einem Datensatz ausschließlich Werte der abhängigen Variablen (also der Y-Variablen) fehlen, können nicht nur bei Gültigkeit der MCAR-Eigenschaft, sondern auch bei Gültigkeit der MAR-Eigenschaft alle Fälle mit fehlenden Werten listenweise aus der Regressionsanalyse ausgeschlossen werden. Denn unter dieser Voraussetzung liefert eine Regressionsschätzung unverzerrte und effiziente Schätzer sowie korrekte Inferenzstatistiken.<sup>19</sup> Eine solche Datenlage gibt es z.B. auch in unserem Antisemitismusdatensatz (s.o.). Alle fehlenden Werte betreffen dort die abhängige Variable „AS\_Index“ (vgl. Tabelle 3). Zudem kann dort auch angenommen werden, dass die Zufälligkeit fehlender Y-Werte bei Kontrolle der Einflüsse von „Alter“ und „Bildung“ bestehen bleibt, sodass eine MAR-Eigenschaft der fehlenden Werte durchaus möglich ist (vgl. Tabelle 6). Aber unabhängig davon, ob die MCAR- oder die MAR-Eigenschaft bei den fehlenden Werten des AS\_Index gilt, kann im Falle von fehlenden Werten, die allein bei der Y-Variablen einer Regressionsschätzung auftreten, davon ausgegangen werden, dass die fehlenden Werte ohne Probleme listenweise aus der Regressionsschätzung ausgeschlossen werden können.

### **(2) Die X-bezogene Voraussetzung**

Wenn bei einer Regressionsanalyse nur die Prädiktoren (also die X-Variablen) fehlende Werte aufweisen und wenn die Wahrscheinlichkeit dafür nicht von der Ausprägung der Y-Variablen (also von der Höhe der Y-Variablen) abhängt, können ebenfalls die Fälle mit fehlenden Werten problemlos listenweise ausgeschlossen werden. Unter dieser Bedingung müssen die betreffenden X-Variablen auch nicht die MAR-Eigenschaft aufweisen. Auch dürfen dann die X-Missings von der wahren (aber unbekannten, weil fehlenden) Ausprägung einer X selbst sowie von anderen X-Variablen abhängen.

---

<sup>18</sup> In der folgenden Auflistung werden bei den Erläuterungen zu einer jeden Anwendungsvoraussetzung die diesbezüglich zitierfähigen Literaturquellen angegeben.

<sup>19</sup> Dazu finden sich formale und empirische Begründungen bei: Allison 2002: 54; Baltès-Götz 2013: 25; Carpenter/Kenward 2013: 24-28; Schafer/Graham 2002: 155.

Wenn also die Wahrscheinlichkeit von X-Missings nicht von Y abhängt, können auch bei einem listenweisen Fallausschluss alle Schätzwerte der Regressionsanalyse als unverzerrt gelten. Und dies ist selbst dann der Fall, wenn die fehlenden X-Werte als MNAR einzustufen sind.<sup>20</sup> In unserem Anwendungsbeispiel (dem Antisemitismusmodell) wäre diese Voraussetzung erfüllt, wenn z.B. der Prädiktor „Bildung“ eine bestimmte Missingquote aufwiese. Wenn diese Missings entscheidend zu einer kumulierten Missingquote von über 5% beitragen würden, könnte man nach der hier referierten Entscheidungsregel die Fälle mit fehlenden Bildungswerten problemlos listenweise aus der Regressionsanalyse ausschließen, weil anzunehmen ist (und evtl. auch empirisch zu überprüfen wäre), dass die fehlenden Werte der Variablen „Bildung“ nicht von den Ausprägungen des AS\_Indexes abhängen dürften.

### **(3) Die Y/X-bezogene Voraussetzung (bei logistischen Regressionen)**

Diese spezielle Voraussetzung für den listenweisen Ausschluss von Fällen gilt nur bei logistischen Regressionsanalysen (vgl. Urban/Mayerl 2011: 323ff). Wenn dort fehlende Werte bei der Y-Variablen und/oder bei der/den X-Variablen vorkommen, und wenn die Wahrscheinlichkeit für diese fehlenden Werte nur von der Y-Variablen abhängt, dann liefert die logistische Regressionsschätzung für fast alle Parameterschätzungen unverzerrte Schätzwerte.<sup>21</sup> Allein der Schätzwert für die Konstante (Intercept) der Regressionsschätzung ist dann verzerrt. Probleme bei einer logistischen Regressionsschätzung mit fehlenden Werten ergeben sich also nur dann, wenn die Wahrscheinlichkeit von Missings sowohl von Y als auch von X abhängt.<sup>22</sup> Hätten wir z.B. in unserem Antisemitismusdatensatz einen binären AS\_Index und wollten deshalb eine logistische Regressionsschätzung mit diesem binären AS-Index als abhängiger Variablen durchführen, könnten wir unter Verweis auf die spezielle Y/X-bezogene Voraussetzung den Einsatz eines listenweisen Fallausschlusses nicht rechtfertigen. Denn es wäre zwar anzunehmen, dass die Wahrscheinlichkeit für fehlende „AS\_Index“-Werte von der Höhe dieser Y-Variablen beeinflusst wird, aber es wäre sicherlich falsch anzunehmen, dass die Wahrscheinlichkeit fehlender Werte bei der X-Variablen „Bildung“ von der Ausprägung des

---

<sup>20</sup> Dazu finden sich formale und empirische Hinweise bei: Allison 2000: 303f; Allison 2002: 6f; Baltes-Götz 2013: 25; Enders 2010: 40; Little/Rubin 2002: 43. Eine diesbezügliche logische Beweisführung liefern Carpenter/ Kenward (dies. 2013: 28f). Eine intuitiv verständliche Erläuterung der Thematik versucht Allison (ders. 2002: 7). Er begründet die Möglichkeit eines listenweisen Ausschlusses von Fällen mit fehlenden Daten bei Vorhandensein der o.g. X-bezogenen Voraussetzungen damit, dass in Samples, die auf der Basis von unabhängigen Variablen (X-Variablen) disproportional geschichtet sind, die Schätzwerte einer Regressionsanalyse nicht verzerrt seien. Und da Samples, in denen die Wahrscheinlichkeit für fehlende Werte ebenfalls nur von den X-Variablen abhängt, als äquivalent zu stratifizierten Samples zu betrachten seien, gelte auch für sie, dass Regressionsschätzungen mit diesen Daten nicht zu verzerrten Resultaten führen müssten.

<sup>21</sup> Es sei hier noch einmal darauf hingewiesen, dass das Gesagte nur für logistische Regressionsschätzungen gilt. Im Falle von OLS-Regressionsschätzungen sollten auch dann, wenn die aufgeführte Voraussetzung No. 3 erfüllt wäre, keine listenweisen Ausschlüsse vorgenommen werden (es sei denn, die Anwendungsvoraussetzungen No. 1 und No. 2 wären erfüllt) und sollten dann die im Folgenden beschriebenen Imputationsverfahren benutzt werden.

<sup>22</sup> Dazu finden sich formale und empirische Begründungen bei: Allison 2002: 7; Carpenter/ Kenward 2013: 30-32.

AS\_Indexes abhängt. Wir könnten dann die Verwendung des listenweisen Fallausschlusses allein über die spezielle Y-bezogene Voraussetzung (s.o.) oder die MCAR-Eigenschaft rechtfertigen.

Im Folgenden wollen wir im Rahmen unseres Anwendungsbeispiels eine Regressionsschätzung des Einflusses der Prädiktoren „Bildung“ und „Alter“ auf das Ausmaß von Antisemitismus (abh. Variable: „AS\_Index“) durchführen und dabei einen listenweisen Ausschluss der Fälle mit fehlenden Werten zulassen. Ein listenweiser Fallausschluss kann bei diesem Beispiel durchgeführt werden, weil von der MCAR-Eigenschaft für die fehlenden Werte des betreffenden Datensatzes auszugehen ist (vgl. SISS-Kapitel 2), und weil (zusätzlich) die spezielle Anwendungsvoraussetzung No. 1 (Y-bezogene Voraussetzung, s.o.) bei den Missings dieses Datensatzes erfüllt ist (Begründung s.o.).

Die folgende SPSS-Syntax 8 steuert eine Regressionsschätzung mit listenweisem Ausschluss aller Fälle mit fehlenden Werten,

#### SPSS-Syntax 8

```
REGRESSION  
/MISSING LISTWISE  
/DEPENDENT AS_Index  
/METHOD=ENTER Alter Bildung.
```

Die folgenden Tabellen 7a und 7b zeigen einige Resultate der Regressionsschätzung mit einem listenweisen Ausschluss aller Fälle mit fehlenden Werten. Aufgrund des listenweisen Ausschlusses werden nicht mehr insgesamt 92, sondern nur noch 77 Fälle für die Regressionsschätzung benutzt. Nach den Ergebnissen dieser Regressionsschätzung kann ca. 10% der Varianz vom AS\_Index durch Einflüsse der beiden Prädiktoren statistisch erklärt werden (Tab. 7a). Die Effekte des Bildungsprädiktors sind hochgradig signifikant, während der Altersprädiktor keinen signifikanten Einfluss auf das Ausmaß von Antisemitismus ausübt (Tab. 7b). Die standardisierten Regressionskoeffizienten  $b^*$  (in SPSS als „Beta“ bezeichnet) von -0,32 und -0.01 verweisen darauf, dass der Bildungseffekt im Vergleich zum Alterseffekt wesentlich stärker ist. Zudem ist der Bildungseffekt negativ ausgerichtet. Mit höherem Bildungsgrad ist also zu erwarten, dass sich das geschätzte Ausmaß von Antisemitismus reduziert.

**Tabelle 7a:** Ergebnisse der Regressionsschätzung nach listenweisem Ausschluss von Fällen mit fehlenden Werten, Teil 1 (Antisemitismusdaten)

Modellübersicht				
Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,331 <sup>a</sup>	,098	,074	1,18874

**Tabelle 7b:** Ergebnisse der Regressionsschätzung nach listenweisem Ausschluss von Fällen mit fehlenden Werten, Teil 2 (Antisemitismusdaten)

Koeffizienten					
Modell	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
	B	Standard-fehler	Beta		
(Konst.)	4,033	,818		6,521	,000
Alter	-,001	,008	-,012	-,103	,918
Bildung	-,315	,114	-,316	-2,764	,007

Die Regressionsschätzung mit 77 Beobachtungsfällen ergibt demnach zusammenfassend die folgende unstandardisierte Regressionsgleichung (vgl. Tabelle 7b):

$$\hat{Y} = 4,033 - 0,001 \text{ Alter} - 0,315 \text{ Bildung} \quad (1)$$

(0,008)      (0,114)

Um einschätzen zu können, inwiefern die Ergebnisse tatsächlich nicht in größerem Ausmaß verzerrt sind, wenn wir „listwise deletion“ bei einem Vorliegen von MCAR einsetzen, haben wir dieselbe Regressionsanalyse mit allen 92 Beobachtungsfällen, d.h. ohne fehlende Werte, durchgeführt (das ist uns nur dadurch möglich, dass wir die fehlenden Werte in unserem Beispieldatensatz künstlich eingebaut haben und die ursprünglichen Werte kennen). Eine solche Schätzung mit allen Fällen („complete cases“) ergibt die folgende Regressionsgleichung:

$$\hat{Y} = 3,855 - 0,003 \text{ Alter} - 0,329 \text{ Bildung} \quad (2)$$

(0,008)      (0,109)

Ein Vergleich der Gleichungen (1) und (2) zeigt, dass keine bedeutsamen substanziellen Unterschiede zwischen den Regressionsergebnissen hinsichtlich Einflussrichtung und Signifikanz der Einflüsse der X-Variablen vorliegen. So ist der Bildungseffekt in beiden Fällen auf dem 5% Signifikanzniveau

statistisch signifikant negativ und der Alterseffekt ist in beiden Schätzungen deutlich nicht signifikant. Auch unterscheiden sich die geschätzten unstandardisierten b-Koeffizienten der beiden X-Variablen in ihrer Höhe nur marginal ( $-0,001$  versus  $-0,003$  und  $-0,315$  versus  $-0,329$ ). Etwas deutlicher variiert der a-Koeffizient, wobei auch dieser Unterschied nicht allzu groß ist ( $4,033$  versus  $3,855$ ).

Zu beachten ist, dass die Ergebnisse dieses Anwendungsbeispiels nicht verallgemeinert werden dürfen, d.h. es kann aus diesem einen Beispiel keine generelle Erkenntnis über die Höhe von Verzerrungen durch listenweisen Ausschluss (listwise deletion) abgeleitet werden. Das Beispiel dient alleine zu Illustrationszwecken. Vgl. dazu auch die Anmerkungen zu Tabelle 11 im Folgenden.

Fassen wir unsere Empfehlungen zum listenweisen Ausschluss von Fällen mit fehlenden Werten noch einmal zusammen:

In der Regressionsanalyse kann bei fehlenden Werten ein listenweiser Ausschluss von Fällen eine gute Wahl sein. Das Verfahren sollte allerdings nur eingesetzt werden,

- a) wenn die fehlenden Werte die MCAR-Eigenschaft erfüllen,

oder:

- b) wenn zumindest eine der drei oben aufgeführten speziellen Voraussetzungen für fehlende Werte bei den Modellvariablen von Regressionsanalysen erfüllt ist,

und:

- c) wenn ein listenweiser Ausschluss nicht zu einem erheblichen Abbau der Fallzahl in einem kleinen Sample führt (mit dann bedeutsamen Konsequenzen für die Teststärke).

Sind diese Voraussetzungen nicht erfüllt, und liegt die kumulierte oder eine variablenspezifische Missingquote oberhalb von 5%, sollte eines der im Folgenden vorgestellten Imputationsverfahren zur Behandlung fehlender Werte eingesetzt werden.

Misstraut man in seiner Regressionsanalyse dennoch den Schätzergebnissen bei listenweisem Fallausschluss, so kann zur Absicherung oder Feststellung der Störanfälligkeit des Verfahrens auch eine Replikation der Schätzung mittels Substitution der fehlenden Werte durch Imputationsverfahren ausgeführt werden (mehr dazu im Folgenden).



### 3.2 Die stochastische Regressionsimputation (sRI)

Mit dem Verfahren der stochastischen Regressionsimputation (sRI) können fehlende Variablenwerte durch geschätzte Werte ersetzt werden. Das gilt aber nur für metrische (d.h. mindestens intervallskalierte) Variablen. Zur Schätzung von fehlenden Werten bei kategorialen Variablen muss das Verfahren der Multiplen Imputation (MI) eingesetzt werden (vgl. dazu das folgende SISS-Kapitel 3.3).

Für die stochastische Regressionsimputation sollten die fehlenden Werte die MAR-Eigenschaft aufweisen. Ein Teil dieser Eigenschaft lässt sich mit einer logistischen Regressionsschätzung überprüfen (wie im SISS-Kapitel 2 mit den Daten zu unserem Antisemitismusmodell gezeigt wird). Da man aber nichts über die Abhängigkeit der fehlenden Werte von ihren tatsächlichen Werten weiß (weil diese Werte unbekannt sind), und weil die Abhängigkeit der Wahrscheinlichkeit fehlender Werte von den Einflüssen anderer Modellvariablen auch bei Gültigkeit der MNAR-Eigenschaft gilt, lässt sich die MAR-Eigenschaft nicht endgültig statistisch überprüfen. Sie muss ergänzend zur statistischen Schätzung der Chance fehlender Werte in Abhängigkeit von Modellprädiktoren (mittels binärer Regression, s.o.) durch zusätzliche analytische Überlegungen hypothetisch bzw. argumentativ begründet werden.

So könnte z.B. bei unseren Antisemitismusdaten vermutet werden, dass bei der Datenerhebung insbesondere Respondenten mit hoher Schulbildung ihre starken Antisemitismuseinstellungen als „sozial unerwünscht“ betrachten und diese deshalb verschweigen. Deshalb wäre die Zufälligkeit fehlender Werte erst bei Kontrolle von Bildungseffekten möglich. Somit wäre plausibel anzunehmen, dass die fehlenden Werte beim AS\_Index die MAR-Eigenschaft aufweisen könnten.<sup>23</sup>

Die Methode der stochastischen Regressionsimputation (sRI) besteht in verfahrenslogischer Hinsicht aus drei Schritten:

**1. Schritt:** Zunächst wird mit den kompletten Fällen eines Datensatzes eine Regressionsschätzung für die Ausprägungen derjenigen Variablen, bei denen ein bedeutsames Ausmaß von fehlenden Werten festgestellt wurde, vorgenommen (egal ob Y oder X-Variable/n). Als Prädiktoren werden dazu alle anderen Modellvariablen benutzt. Auch die abhängige Variable des Regressionsmodells sollte als Prädiktor spezifiziert werden, wenn es um die fehlenden Werte bei einer X-Variablen geht (vgl.

---

<sup>23</sup> Wie in SISS-Kapitel 3.1 erläutert, können wir die fehlenden Werte beim AS\_Index mit einem listenweisen Fallausschluss aus der Analyse herausnehmen, wenn für sie die spezielle Y-bezogene Eigenschaft gilt und wenn zudem die Fallzahl auch nach diesem Ausschluss noch ausreichend hoch ist (vgl. Kap. 3.1).

dazu Allison 2000: 307). Als ergänzende Prädiktoren können Interaktionsvariablen verwendet werden (auch wenn diese keine Modellvariablen sind). Zusätzlich können in die Schätzung sog. „Hilfsvariablen“ als Prädiktoren einbezogen werden. Das sind Variablen, die zwar im zu analysierenden Regressionsmodell nicht vertreten sind, die aber dennoch für die Prognose von Werten der betroffenen abhängigen Variablen wichtig sein können (in unserem Antisemitismusbeispiel wird das z.B. nachfolgend die metrische Variable „Links/Rechts-Orientierung“ sein).

**2. Schritt:** Mit dem Schätzergebnis aus Schritt 1 wird eine lineare Schätzgleichung zur Prognose der Werte, die im nicht reduzierten Datensatz fehlen, erstellt. In die Schätzgleichung werden die beobachteten/erfragten Werte von Respondenten eingegeben, die zwar bei einer bestimmten Frage ihre Antwort verweigert haben, die aber ansonsten ausreichende empirische Informationen zu den Prädiktoren der jeweiligen Schätzgleichung geliefert haben. Auf diese Weise können für jede dieser Personen die fehlenden Werte der betreffenden Variablen unter Verwendung dieser Schätzgleichung prognostiziert werden.

**3. Schritt:** Die in Schritt 2 geschätzten Variablenwerte werden anstelle der fehlenden Werte in den Datensatz eingespeist (imputiert). Bei jedem Untersuchungsfall (Person, Organisation o.ä.) mit einem fehlenden Wert auf der Kriteriumsvariablen der Regressionsschätzung (Schritt 2) wird die Leerstelle dieses fehlenden Wertes durch einen prognostizierten Wert ersetzt.

**Zusätzlich** wird bei der *stochastischen* Regressionsimputation im Unterschied zur traditionellen Regressionsimputation jeder einzelne Schätzwert durch Hinzufügung einer Zufallskomponente modifiziert. Die stochastische Erweiterung der Regressionsschätzung erfolgt dabei dadurch, dass den einzelnen prognostizierten Werten (Schätzwerten) zufällig ausgewählte Residualwerte hinzu addiert oder von diesen subtrahiert werden. Diese Werte stammen aus der Menge aller Residuen, die bei der Regressionsschätzung derjenigen Variablen, die die fehlenden Werte aufweist (in unserem Beispiel: AS\_Index), unter Verwendung aller kompletten Fällen anfallen (im Schritt 1, s.o.). Die Residuenwerte werden dabei aus der hypothetisch als normalverteilt angenommenen Residuenmenge (mit einem Mittelwert von 0 und einer Varianz, die der beobachteten Residuenvarianz entspricht) zufällig ausgewählt.

Durch Verwendung der Zufallskomponente wird eine Überanpassung der neu geschätzten, ehemaligen Missings an die Ausprägungen ihrer Prädiktoren vermieden und die geschätzten Werte hängen

nicht mehr allein von den empirischen Werten der X-Variablen in der Schätzgleichung ab (im Schritt 2, s.o.).

Ohne die stochastische Ergänzung der Regressionsschätzung für jeden einzelnen fehlenden Wert hätten die Parameterschätzungen insbesondere bei hohen Missingquoten eine stark reduzierte Streuung und würden verzerrte Werte aufweisen (auch unterschätzte Standardfehler). Wenn jedoch die fehlenden Werte den MAR-Bedingungen entsprechen und wenn gleichzeitig die Prognose fehlender Werte um eine Zufallskomponente erweitert wird, sind die Schätzer der sRI unverzerrt. Allerdings weisen die mittels sRI imputierten Werte keine statistischen „Unsicherheiten“ auf, was bedeutet, dass die geschätzten Werte im statistischen Sinne wie empirische Werte behandelt werden. Das kann vor allem bei inferenzstatistischen Schlüssen von Nachteil sein, da durch die fehlende Unsicherheit in den ersetzten Werten oftmals zu kleine Standardfehler ermittelt werden, so dass Effekte zu schnell signifikant werden können. Ein Verfahren, das diese Unsicherheiten berücksichtigt, ist die Multiple Imputation (s.u.). So fallen bei der MI auch die Standardfehler in aller Regel größer aus als bei der sRI, sodass die MI konservativere inferenzstatistische Schlüsse erzeugt.

In der Methodenforschung gibt es konträre Auffassungen darüber, ob auch fehlende Werte bei der abhängigen Y-Variablen eines Regressionsmodells mittels sRI substituiert werden sollten. Die Kritiker eines solchen Vorgehens verweisen dabei auf einen möglichen Zirkelschluss, der dadurch entsteht, dass bei der sRI zunächst mit den X-Prädiktoren die fehlenden Y-Werte geschätzt werden und sodann mit denselben X-Prädiktoren die Abhängigkeitsstruktur derselben Y-Variablen (nunmehr ergänzt um die Fälle mit substituierten fehlenden Werten) erforscht werden soll.<sup>24</sup> Die Befürworter eines solchen Verfahrens, zu denen wir auch gehören, verweisen in diesem Zusammenhang darauf, dass sich durch die Verwendung von Hilfsvariablen und Zufallskomponenten bei der Prognose fehlender Werte die beiden Schätzungen deutlich voneinander unterscheiden. Zudem kann darauf hingewiesen werden, dass die Prognose fehlender Werte mittels X-Prädiktoren nicht willkürlich ist, sondern die empirisch gegebene Abhängigkeitsstruktur zwischen Y- und X-Variablen nutzt.<sup>25</sup>

Im Folgenden wollen wir eine sRI für die fehlenden Werte beim AS\_Index mit unseren Antisemitismusdaten durchführen. Als Prädiktoren zur Schätzung der erwarteten Werte, welche die fehlenden Werte beim AS\_Index ersetzen sollen, benutzen wir die Modellvariablen „Alter“ und „Bildung“ sowie die ergänzende Hilfsvariable „Links/Rechts-Orientierung“. Die diesbezüglichen SPSS-Anweisungen finden sich in der folgenden SPSS-Syntax 9.

---

<sup>24</sup> Vgl. Cohen et al. 2003: 446.

<sup>25</sup> Vgl. Acock 2005: 1026; Allison 2002: 54; Baltes-Götz 2013:33; Enders 2010: 48; Little/ Rubin 2002: 66.

### SPSS-Syntax 9

```
MVA VARIABLES = AS_Index Alter Bildung Linksrechts  
/REGRESSION AS_Index WITH Alter Bildung Linksrechts  
(ADDDTYPE=NORMAL OUTFILE='AS2.sav').
```

Mit dem neu gebildeten AS-Datensatz „AS2.sav“, in dem sich nun auch die neue abhängige Variable „AS\_Index“ befindet (die keine fehlenden Werte mehr aufweist), replizieren wir nunmehr die Regressionschätzung, die wir im SISS- Kapitel 3.1 auch schon unter Verwendung von „listwise deletion“ durchgeführt haben:

### SPSS-Syntax 10

```
REGRESSION  
/DEPENDENT AS_Index  
/METHOD=ENTER Alter Bildung.
```

Die folgende Tabelle 8 ermöglicht einen Vergleich der Schätzergebnisse der beiden Regressionsanalysen, von denen eine Analyse unter Verwendung von „listwise deletion“ (Modell 1) und die andere Analyse unter Verwendung von „stochastischer Regressionsimputation“ (Modell 2) durchgeführt wurde. Zudem zeigt Tabelle 8 auch noch die Ergebnisse, die bei der Regressionschätzung mit den kompletten Daten (ohne Missings) erzielt wurden (Modell 0).

Bei der ersten Regressionsanalyse mit fehlenden Werten (Modell 1) wurden die Fälle mit Missings bei der abh. Variablen „AS\_Index“ mittels „listwise deletion“ ausgeschlossen. Um dieses Verfahren zu rechtfertigen wurde angenommen, dass die spezielle „Y-bezogene Voraussetzung“ (vgl. SISS-Kapitel 3.1) für die analysierten Daten zutrifft. Auch kann zur Rechtfertigung dieser Methode die MCAR-Eigenschaft der fehlenden Werte benutzt werden. Durch den listenweisen Ausschluss von Fällen mit fehlenden Werten standen für diese Regressionsanalyse nur noch insgesamt 77 von 92 Fällen zur Verfügung (vgl. Tabelle 8).

Bei der anderen Regressionsanalyse mit fehlenden Werten (Modell 2) wurden die Missings der abhängigen Variablen „AS\_Index“ mittels stochastischer Regressionsimputation substituiert. Zur Begründung wurde eine MAR-Eigenschaft der fehlenden Werte postuliert. Diese Regressionsanalyse konnte aufgrund des gewählten Imputationsverfahrens alle 92 Fälle für die Schätzung verwenden.

Ein Vergleich der Schätz- und Testwerte aus „Modell 1“ bzw. aus „Modell 2“ mit den Ergebnissen der „wahren“ Regressionsschätzung „Modell 0“ zeigt, dass die Werte zwar geringfügig bis mittelstark voneinander abweichen, dass aber die Abweichungen an keiner Stelle so stark sind, dass sie zu unterschiedlichen substanziellen Interpretationen der Schätzergebnisse führen würden (vgl. Tabelle 8). Insbesondere die Schätzwerte zum signifikanten Effekt des Prädiktors „Bildung“ (der Alterseffekt ist nicht signifikant) führen mit unseren Antisemitismusdaten zu sehr ähnlichen Ergebnissen.

**Tabelle 8:** Ergebnisse einer Regressionsschätzung (Modell 0) ohne Missings (nur „complete cases“) und von zwei Schätzungen mit unterschiedlichen Verfahren zur Behandlung fehlender Werte (Modelle 1 u. 2) (Antisemitismusmodell)

	Modell 0	Modell 1	Modell 2
	„complete cases“	„listwise deletion“	„stoch. Regr.-imputation“
Brutto-N	92	92	92
Missingquote	---	16,3%	vor Imputation: 16,3%; nach Imputation: 0,0%
Netto-N	92	77	92
a			
Konstante	3,855	4,033	4,309
b			
Alter	0,003	-0,001	-0,007
Bildung	-0,329	-0,315	-0,302
SE			
Konstante	0,610	0,618	0,604
Alter	0,008	0,008	0,008
Bildung	0,109	0,114	0,108
b*			
Alter	0,045	-0,012	-0,101
Bildung	-0,322	-0,316	-0,304
Signifikanz (p-Wert)			
Konstante	0,000	0,000	0,000
Alter	0,670	0,918	0,357
Bildung	0,003	0,007	0,006
R-Quadrat	0,116	0,098	0,081

Zusammenfassend können wir für das sRI-Verfahren in der Regressionsanalyse einige Vorteile (+) und auch einige Nachteile (–) erkennen:

- (+) Die sRI ist zumindest für Anwender von Regressionsanalysen ein intuitiv verständliches Verfahren zur Substitution fehlender Werte.
- (+) Datensätze, bei denen fehlende Werte durch sRI substituiert wurden, können mit beliebigen statistischen Methoden ausgewertet werden. Die Datensätze können auch zur Auswertung mit anderen Statistik-Programmen problemlos exportiert werden.
- (+) Die Parameterschätzungen in einer stochastischen Regressionsanalyse sind bei zutreffenden MAR-Annahmen im Prinzip unverzerrt.

- (+) Bei kleinem N bleibt die Fallzahl erhalten bzw. wird durch das sRI-Verfahren nicht weiter reduziert.
- (-) Es können nur fehlende Werte von metrischen, mindestens intervallskalierten Variablen substituiert werden (als Prädiktoren zur Schätzung fehlender Werte können jedoch dichotome und metrische Variablen benutzt werden).
- (-) Es müssen für eine robuste sRI prognosestarke Prädiktoren und Hilfsvariablen zur Verfügung stehen.
- (-) Den mittels sRI imputierten Werten wird in der Regressionsanalyse die gleiche Bedeutung gegeben wie den empirischen Befragungs-/Beobachtungswerten. Die damit verbundene Unsicherheit bzw. das damit verbundene Risiko bleibt unberücksichtigt und somit können die mit sRI imputierten Werte die beobachteten Werte niemals gänzlich unverzerrt ersetzen.<sup>26</sup>

### 3.3 Die Multiple Imputation (MI)

Die Multiple Imputation (MI) ist z.Z. eines der fortschrittlichsten statistischen Verfahren zur Behandlung fehlender Werte in der Regressionsanalyse und hat einige Vorteile im Vergleich zur stochastischen Regressionsimputation (sRI). So kann das Verfahren z.B. fehlende Werte sowohl bei metrischen Variablen (mindestens intervallskaliert) als auch bei kategorialen Variablen (bi- oder multinomial, ungeordnet oder geordnet) substituieren. Zwar setzt die MI formallogisch voraus, dass zumindest ein MAR-Ausfallmechanismus die fehlenden Werte erzeugt hat. Sie kann aber auch unter MCAR-Bedingungen sehr gute Ergebnisse liefern.<sup>27</sup>

Die Verfahrenslogik der MI besteht darin, dass in diesem Verfahren jeder fehlende Wert nicht nur durch einen einzigen, sondern durch mehrere simulierte Werte ersetzt wird. Durch mehrfache Simulation ergeben sich mehrere Datensätze mit unterschiedlich imputierten Werten. Jeder einzelne Datensatz kann dann mit der vom Forscher gewünschten statistischen Analysemethode so ausgewertet werden (in unserem Fall ist das die Regressionsanalyse) als wäre er der komplette Datensatz. Natürlich sind die mit jedem Datensatz und den darin imputierten Werten geschätzten Regressionsresultate nicht alle identisch, sondern variieren untereinander. Deshalb werden die einzelnen Regressionsschätzungen beim MI-Verfahren nach besonderen Regeln (den sog. „Rubins Regeln“, s.u.) zusammengefasst. Es ergeben sich dann durchschnittliche Parameterschätzwerte und durchschnittliche Standardfehler. Zugleich wird bei dieser Zusammenführung die Unsicherheit der zahlreichen Imputationen

---

<sup>26</sup> Carpenter/Kenward 2013: 38.

<sup>27</sup> Nach den Simulationsergebnissen von Schafer/Graham 2002: 169.

durch einen angepassten Schätzwert bei der Kalkulation der durchschnittlichen Standardfehler berücksichtigt.

Am einfachsten lässt sich die Verfahrenslogik der Multiplen Imputation als eine Abfolge von drei Schritten verstehen (vgl. dazu auch die folgende Abbildung 1). Im Folgenden werden die allgemeinen Prinzipien dieser drei Schritte kurz erläutert:

Schritt 1 (Imputationsphase): Es werden mehrere Kopien des Datensatzes (Daumenregel:  $m=20$  Datensätze, s.u.) geschaffen, in die jeweils unterschiedliche Schätzwerte für die fehlenden Werte imputiert werden. Die fehlenden Werte werden mit Hilfe von Regressionsmodellen (unter Verwendung von metrischen und binären Prädiktoren sowie von Hilfsvariablen) geschätzt. Die Schätzungen erfolgen bei fehlenden metrischen Variablenwerten mittels OLS-Regression und bei fehlenden kategorialen Werten mittels logistischer Regression, wobei natürlich auch stets die hier in SISS-Kapitel 3.2 vorgestellten Zufallskomponenten in die Schätzung aufgenommen werden. Dieses in SPSS implementierte Verfahren wird als FCS-Methode bezeichnet (*Fully Conditional Specification Methode*).<sup>28</sup> Die Methode wird im Anhang des vorliegenden SISS-Berichts ausführlich vorgestellt.<sup>29</sup>

Schritt 2 (Analysephase): Mit jedem einzelnen Datensatz wird die gewünschte Datenanalyse (hier: Regressionsanalyse) durchgeführt. Dabei werden so viele unterschiedliche Schätzwerte für jeden einzelnen Modellparameter ermittelt, wie zu Beginn des Imputationsverfahrens als „m“ festgelegt wurden.

Schritt 3 (Poolingphase): Sind in der Analysephase (s.o.) m-verschiedene Regressionsschätzungen durchgeführt worden, müssen im Rahmen einer Poolingphase die in jedem Modell geschätzten Parameter zu durchschnittlichen Parameterschätzwerten zusammengeführt werden. Dazu gibt es für

---

<sup>28</sup> Die Schätzmethode, welche zur Imputation der fehlenden Werte herangezogen wird (in SPSS ist das die FCS-Methode), ist jeweils vom gewählten Statistik-Softwarepaket abhängig, wohingegen die Verfahrenslogik der MI in den meisten Statistik-Paketen stets die gleiche ist. Die FCS-Methode wird im Anhang zu dieser Schrift ausführlich vorgestellt. Anzu-merken ist, dass die in anderen Statistik-Softwarepaketen verwendeten Verfahren teilweise auch andere Voraussetzungen an die Daten stellen. Bspw. gibt es Verfahren die eine multivariate Normalverteilung aller Daten voraussetzen, sodass damit nur metrische Variablen imputiert werden könnten. Das hier vorgestellte FCS-Verfahren ist in dieser Hinsicht anderen Verfahren überlegen, da es keine solche Annahme (explizit) trifft (vgl. van Buuren 2007: 227).

<sup>29</sup> Die FCS-Methode, welche die zu imputierenden Werte u.a. mittels Zufallsziehungen bestimmt, basiert auf Konzepten der Bayesschen Statistik. Sie benutzt eine iterative Monte-Carlo-Simulationstechnik (MCMC-Methode = Markov Chain Monte Carlo-Methode). Dabei werden zuerst aus einer posteriori Verteilung (auf Grundlage der empirischen Beobachtungen) bestimmte Parameter gezogen, bevor aus einer weiteren posteriori Verteilung (mit Hilfe der emp. Beobachtungen und den gezogenen Parametern) die schlussendlichen Werte für die jeweiligen Missings hervorgehen. Vgl. dazu die Erläuterungen im Anhang des vorliegenden SISS-Berichts. Dort wird die FCS-Methode stark vereinfacht ohne Bayes-Modellierung vorgestellt. Verstehbare Einführungen in die Bayesschen Grundlagen und Techniken der MI gibt es in: Allison 2002: 27-88; Enders 2010: 187-286.

einige Parameter (z.B. für den unstandardisierten Regressionskoeffizienten) bestimmte Regeln (die sog. „Rubins Regeln“, s.u.). Einige der derart ermittelten, durchschnittlichen Schätzwerte berücksichtigen auch die Unsicherheit, die mit der Schätzung fehlender Werte mittels MI verbunden ist. So sind üblicherweise die durchschnittlichen Standardfehler der geschätzten Regressionskoeffizienten bei Analysen mit vorausgegangener MI größer als bei vergleichbaren Regressionsanalysen mit anderen Imputationsverfahren. Dadurch kann es für diese Koeffizienten auch schwieriger werden, eine statistische Signifikanz zu erreichen.

Bei der Durchführung eines MI-Verfahrens ist Folgendes zu beachten:

- a) Es können im MI-Verfahren die Missings von metrischen und kategorialen Variablen ersetzt werden. Für die Schätzung der zu imputierenden Werte sind metrische, binomiale oder kategoriale Prädiktoren bzw. Hilfsvariablen erforderlich, wobei SPSS im MI-Verfahren automatisch alle kategorialen Variablen in binomiale Dummy-Variablen rekodiert. Als Prädiktoren sollten alle Variablen des Regressionsmodells eingesetzt werden (dazu gehört auch diejenige Variable deren fehlende Werte zu schätzen sind<sup>30</sup>). Als Hilfsvariablen werden Prädiktoren bezeichnet, die nicht im zu analysierenden Regressionsmodell vertreten sind, die aber das Antwortverhalten bei Messung des Items mit den fehlenden Werten beeinflussen könnten. Je mehr Prädiktoren und Hilfsvariablen eingesetzt werden können, umso besser ist dies für die Schätzung fehlender Werte.
- b) Für die Festlegung der Anzahl simulierter Teilstichproben (m) gilt grundsätzlich: 1) Je mehr umso besser, und: 2) Je höher die Missingquote umso höher sollte m sein. So wird z.B. oftmals empfohlen, so viele Imputationen, wie Prozent-Missings vorhanden sind, anzufordern. Allerdings reicht nach unseren Erfahrungen ein m von 20 absolut aus, um auch bei hohen Missingquoten robuste Parameterschätzwerte und brauchbare Teststatistiken zu erhalten.<sup>31</sup>

<sup>30</sup> Dies ist ein Vorteil der iterativen FCS-Methode (*Fully Conditional Specification Methode*), die von SPSS im MI-Verfahren eingesetzt wird (und die wir im Anhang dieses SISS-Beitrags ausführlicher erläutern). Bei dieser Methode werden auch alle Variablen, deren Missings ersetzt werden sollen, als Prädiktoren benutzt, sodass die gesamte beobachtete Information eines Datensatzes im Schätzverfahren genutzt werden kann. Ansonsten würde das self-learning-Prinzip des darin implementierten Algorithmus geschwächt, da bestimmte empirische Informationen nicht zur Verfügung ständen. Denn wenn diejenige Variable, die im FCS-Verfahren ersetzt wird, nicht als Prädiktor dient, wird für die zweite zu ersetzende Variable diese erste ersetzte Variable nicht herangezogen. Wenn also gilt: „ $X_1 = X_2, X_3 \dots X_k$ “ und  $X_1$  würde nicht als Prädiktor benutzt, dann wäre: „ $X_2 = X_3 \dots X_k$ “ ohne Berücksichtigung von  $X_1$ . Mit  $X_1$  als Prädiktor gilt im FCS-Verfahren stattdessen: „ $X_2 = X_3 \dots X_k, X_{1+t}$ “ (vgl. dazu den Anhang). Dementsprechend sollte grundsätzlich in einem multiplen Imputationsmodell nicht zwischen unabh. und abh. Variablen unterschieden werden (vgl. dazu auch Enders 2010: 202).

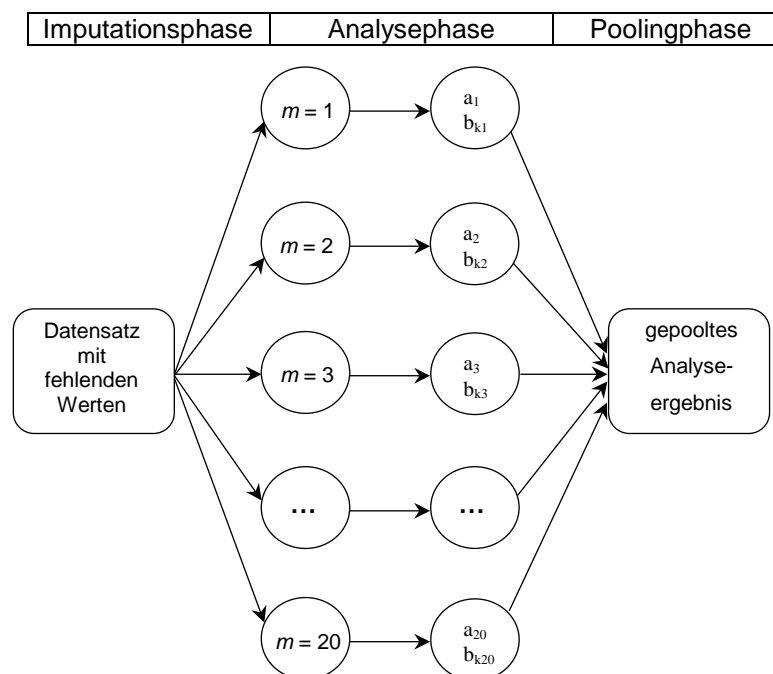
<sup>31</sup> Ab einer Missingquote von ca. 50% kann die MI auch mit  $m=40$  durchgeführt werden (Graham 2009: 561). Allerdings zeigte ein brauchbares Simulationsbeispiel, dass auch mit  $m=20$  und einer hohen Missingquote von 80% bei 1000 Samples (jeweils mit  $N=50$ ) unter MCAR- und MAR-Bedingungen die "Coverage" im MI-Verfahren bei 95% lag. Dabei wird als "Coverage" der prozentuale Anteil aller 1000 Konfidenzintervalle bezeichnet, die den wahren Regressionskoeffizienten, der im MI-Verfahren geschätzt werden musste, enthalten (Schafer/Graham 2002: 169).



Dazu muss bei Durchführung einer MI in SPSS die dort enthaltene Voreinstellung von  $m=5$  auf  $m=20$  hochgesetzt werden (s.u.).

- c) Die Anzahl der im MI-Verfahren auszuführenden Imputationen ( $m$ ) darf nicht mit der maximalen Anzahl an Iterationen ( $M=\text{moves}$ ) verwechselt werden, die im iterativen Monte-Carlo-Simulationsverfahren (s.o.) ausgeführt werden müssen, um zu einem stabilen Schätzergebnis zu gelangen. In der MI von SPSS ist eine maximale Anzahl von  $M=10$  voreingestellt. Erst wenn 10 Moves nicht ausreichen, um das Schätzverfahren mit einer stabilen Schätzung abzuschließen (was SPSS durch eine Fehlermeldung mitteilen würde), sollte dieser Wert erhöht werden.

**Abbildung 1:** Die Logik der Multiplen Imputation am Beispiel einer Regressionsanalyse (nach Enders 2010:188)



Im Folgenden wird gezeigt, in welcher Weise das SPSS-MI-Verfahren eingesetzt werden kann,

- um in unserem Antisemitismusdatensatz die fehlenden Werte beim Antisemitismusindex „AS\_Index“ durch geschätzte Werte zu ersetzen (erster MI-Schritt),
- um mit den  $m$  geschätzten Datensätzen die Regressionsschätzungen für die Prädiktoreffekte von „Bildung“ und „Alter“ zu erhalten (zweiter MI-Schritt),
- um sodann die  $m$  Regressionsschätzungen zu einem gepoolten Gesamtergebnis zusammenzuführen (dritter MI-Schritt).

Als zusätzliche Hilfsvariablen werden bei unserem MI-Verfahren die binäre Variable „Geschlecht“ und die metrische Variable „Links/Rechts-Orientierung“ benutzt. Die folgende SPSS-Syntax 11.1 enthält die Anweisungen für eine Regressionsanalyse mit vorausgehendem MI-Verfahren zur Ersetzung der fehlenden Werte nach der FCS-Methode, wobei mit dieser Syntax nicht nur die fehlenden Werte in der Variablen „AS\_Index“ sondern auch in allen anderen Modellvariablen ersetzt werden (sollte nur eine Imputation der fehlenden Werte in „AS\_Index“ gewünscht werden, so ist dafür die folgende SPSS-Syntax 11.2 zu benutzen, mit der allerdings entsprechend der FCS-Logik auch die Variable „AS\_Index“ als Prädiktor in das MI-Verfahren einbezogen wird).

### SPSS-Syntax 11.1

```

DATASET DECLARE AS_imputiert WINDOW=HIDDEN.
MULTIPLE IMPUTATION AS_Index Alter Bildung Linksrechts Geschl
/IMPUTE METHOD=FCS NIMPUTATIONS=20
/IMPUTATIONSUMMARIES MODELS
/OUTFILE IMPUTATIONS=AS_imputiert.

DATASET ACTIVATE AS_imputiert WINDOW=ASIS.
REGRESSION
/STATISTICS COEFF OUTS CI(95) R ANOVA
/DEPENDENT AS_Index
/METHOD=ENTER Alter Bildung.

```

### SPSS-Syntax 11.2

```

DATASET DECLARE AS_imputiert WINDOW=HIDDEN.
MULTIPLE IMPUTATION AS_Index Alter Bildung Linksrechts Geschl
/IMPUTE METHOD=FCS NIMPUTATIONS=20
/CONSTRAINTS Alter ( ROLE=IND)
/CONSTRAINTS Bildung ( ROLE=IND)
/CONSTRAINTS Linksrechts ( ROLE=IND)
/CONSTRAINTS Geschl ( ROLE=IND)
/IMPUTATIONSUMMARIES MODELS
/OUTFILE IMPUTATIONS=AS_imputiert.

DATASET ACTIVATE AS_imputiert WINDOW=ASIS.
REGRESSION
/STATISTICS COEFF OUTS CI(95) R ANOVA
/DEPENDENT AS_Index
/METHOD=ENTER Alter Bildung.

```

Die Anweisung „NIMPUTATIONS=20“ steht in SPSS-Syntax 11.1 und 11.2 für die Anzahl an zu erzeugenden Datensätzen „m“. In SPSS-Syntax 11.2 werden mit „ROLE=IND“ alle Indikatoren gekennzeichnet, die im MI-Verfahren zur Schätzung der fehlenden Werte eingesetzt werden und deren evtl. fehlende Werte nicht auch noch zusätzlich imputiert werden sollen (wie es nach SPSS-Syntax 11.1 erfolgt). Wichtig ist zudem, dass im SPSS-Datensatz für alle Variablen das Messniveau korrekt spezifiziert wurde (evtl. mit SPSS-Syntax 14).

Das in SPSS zusammengeführte Ergebnis der zwanzig Regressionsschätzungen (m=20), die mit der SPSS-Syntax 11.1 erzeugt wurden, zeigen die folgenden Tabellen 9a/b. Am Ende dieses SISS-

Kapitels werden wir diese Schätzung noch mit den Ergebnissen der Regressionsschätzung unter Verwendung einer stochastischen Regressionsimputation und unter Verwendung eines listenweisen Ausschlusses fehlender Werte vergleichen (Tab.11). Zunächst werden wir jedoch die einzelnen Statistiken in den Tabellen 9a und 9b kurz erläutern.

Tabelle 9a: Zusammengefasster Regressionsschätzung nach MI-Verfahren (Antisemitismusdaten), Teil 1 der bearbeiteten SPSS-Ausgabe

		Nicht standardisierte Koeffizienten		Standard. Koef.	t	Sig.
		B	Standardfehler	Beta		
zusammengefasst	Konst.	4,069	,625		6,539	,000
	Alter	-,001	,008		-,218	,828
	Bildung	-,315	,116		-2,837	,005

Tabelle 9b: Zusammengefasster Regressionsschätzung nach MI-Verfahren (Antisemitismusdaten), Teil 2 der bearbeiteten SPSS-Ausgabe

		Konfidenzintervall für B (95,0%)		Anteil fehl. Info.	Relative Zunahmevarianz	Relative Effizienz
		Untergrenze	Obergrenze			
zusammengefasst	Konst.	2,841	5,274	,065	,069	,997
	Alter	-,018	,014	,101	,111	,995
	Bildung	-,534	-,098	,077	,083	,996

Die Tabellen 9a und 9b enthalten die zusammengeführten Ergebnisse von 20 Regressionsanalysen mit 20 verschiedenen Datensätzen, in denen die fehlenden Werte jeweils durch andere, mittels MI-Verfahren geschätzte Werte ersetzt worden sind. Diese zusammengeführten Ergebnisse werden von SPSS erst in der letzten Zeile einer langen Tabelle aufgelistet (die hier nicht abgedruckt wird). In dieser Liste werden die einzelnen Ergebnisse aller m Regressionsschätzungen ausgegeben (in unserem Beispiel ist m=20 und folglich werden in SPSS die Ergebnisse von 20 Regressionsschätzungen ausgegeben; danach erst wird das Ergebnis der kombinierten Regressionsschätzung berichtet). Im Einzelnen werden von SPSS folgende kombinierten Kennwerte berechnet:

- **Nicht standardisierte Regressionskoeffizienten „B“**

Diese Koeffizienten werden von uns im vorliegenden Text häufig als b-Koeffizienten bezeichnet. Sie sind die arithmetischen Mittelwerte der nicht-standardisierten Regressionskoeffizienten aus allen m (hier m=20) Regressionsschätzungen, die im Rahmen des MI-Verfahrens durchgeführt wurden.

- **Standardfehler (SE)**

Die Berechnung der einzelnen Standardfehler für alle zusammengeführten Einzelschätzungen berücksichtigt die Unsicherheit, die mit Regressionsanalysen unter Verwendung imputierter Werte verbunden ist. Zur Kalkulation des zusammengeführten Standardfehlers wird die Varianz innerhalb („within“) der einzelnen Parameterschätzungen ( $V_W$ ) und die Varianz zwischen („between“) den einzelnen Parameterschätzungen ( $V_B$ ) benötigt.<sup>32</sup> Diese ergeben sich:

bei  $V_W$ : als Durchschnitt der quadrierten SEs bei Schätzung jedes einzelnen Modellparameters in allen  $m$  Datensätzen;

bei  $V_B$ : als Varianz der  $m$  Parameterschätzungen eines jeden Modellparameters.

Der kombinierte Standardfehler für jeden einzelnen Modellparameter wird berechnet als:

$$SE = \sqrt{V_W + (1 + 1/m) \times V_B} \quad (3)$$

- **Standardisierte Regressionskoeffizienten (in SPSS „Beta“ genannt)**

Die standardisierten Regressionskoeffizienten (im vorliegenden Text von uns auch als  $b^*$  bezeichnet) werden von SPSS an dieser Stelle nicht ausgegeben und müssen separat berechnet werden (s.u.).

- **Teststatistik „t“**

Die Testwerte „t“ ergeben sich an dieser Stelle als Quotient aus dem zusammengeführten unstandardisierten Regressionskoeffizienten „b“ (s.o.) und den zusammengeführten Standardfehlern „SE“ (s.o.):

$$t = b / SE \quad (4)$$

Die Freiheitsgrade (df) für die dazugehörigen Signifikanztests werden berechnet als:

$$df = (m-1) [1 + M(SE) / (1 + 1/m)V_B]^2 \quad (5)$$

mit  $M(SE)$  als arithmetischer Mittelwert aller  $m$  dazugehörigen Standardfehler.

- **Signifikanz „Sign.“**

Berichtet werden die „Signifikanzen“, d.h. die empirischen Irrtumswahrscheinlichkeiten (p-Werte), der Testergebnisse mit der jeweiligen Teststatistik „t“ (s.o.). Werte kleiner als 0,05 geben an, dass der Koeffizient bei einem Signifikanzniveau von 5% „statistisch signifikant“ ist.

<sup>32</sup> Die folgenden Formeln 3 bis 5 orientieren sich an den Notationen in Fox 2016: 623 und Graham/Hofer 2000: 211.

- **Konfidenzintervall für B (95,0%)**

Berichtet wird die Breite der 95%igen Konfidenzintervalle, die sich für jeden kombinierten Parameterschätzwert (s.o.) mit dem jeweiligen kombinierten Standardfehler (s.o.) berechnen lassen.

- **Anteil fehlender Information (AFI)**

Berichtet werden die Anteile fehlender Informationen (AFI) bei den einzelnen Parameterschätzungen. Diese Zahlen sind nicht identisch mit den einzelnen Missingquoten. Vielmehr quantifizieren sie den Einfluss fehlender Werte auf die Varianz jeder einzelnen Parameterschätzung. Der jeweilige Anteil fehlender Werte (AFI) kann bei einer sehr hohen Anzahl von Imputationen berechnet werden nach:<sup>33</sup>

$$AFI = (V_B + V_B / m) / V_T \quad (6)$$

So bedeutet ein Wert von 0,077, dass sich in der zusammengeführten Regressionsschätzung insgesamt 7,7% der Totalvarianz des Regressionskoeffizienten von „Bildung“ ( $V_T$ ) durch die Imputation fehlende Werte ergibt.<sup>34</sup> Die AFI-Werte sollten möglichst klein sein und auf jeden Fall unterhalb von 0,25 liegen. Ist ein AFI-Wert sehr hoch, sollte die Anzahl der Imputationen ( $m$ ) deutlich heraufgesetzt werden.

- **Relative Zunahmevarianz (RZV)**

Ähnlich wie der AFI (s.o.) ist auch die relative Zunahmevarianz (RZV) ein Messwert für den Einfluss der Imputation auf die Varianz der Parameterschätzwerte aus den  $m$  Regressionsschätzungen. Sie kann berechnet werden nach:

$$RZV = AFI / (1 - AFI) \quad (7)$$

Der RZV-Wert informiert über die relative Veränderung der Varianz der Parameterschätzwerte durch Imputation fehlender Werte. Ein Wert von 0,111 bedeutet, dass die Parametervarianz mit fehlenden Werten um 11,1% größer ist als die Parametervarianz bei einer Schätzung ohne fehlende Werte. Wenn die fehlenden Werte keinen Einfluss auf die Varianz eines bestimmten Parameters hätten, wäre die RZV gleich 0,0.

<sup>33</sup> Vgl. zu Gl.(6) bis Gl.(8) Enders 2010: 204f, 212f, 214, 225f.

<sup>34</sup> Überraschenderweise kann der AFI bei Variablen, die im Datensatz keine Missings aufweisen (wie auch hier bei „Alter“ und „Bildung“), dennoch Werte von größer 0 annehmen. Dies liegt vor allem daran, dass sich die Anteile aus Varianzschätzungen und nicht aus empirischen Missingquoten ableiten. Zudem können durch Korrelation einer Variablen mit einer anderen Modellvariablen, die eine hohe Missingquote aufweist, hohe AFI-Werte entstehen. Weitere Informationen dazu finden sich in Allison 2002: 48f.

- **Relative Effizienz (RE)**

Die relative Effizienz (RE) ist ein Maß für die Qualität einer statistischen Analyse, bei der ein MI-Verfahren eingesetzt wurde. Die RE lässt sich so verstehen, dass sie über einen Vergleich einer bestimmten MI-gestützten Analyse mit einer zweiten Analyse, bei der nicht nur  $m$  simulierte Datensätze, sondern unendlich viele simulierte Datensätze eingesetzt werden, informiert. Der RE-Wert berechnet sich nach:

$$RE = (1 + AFI/m)^{-1} \quad (8)$$

Ein RE-Wert von 0,997 bedeutet, dass die Parametervarianz einer Schätzung mit unendlich vielen simulierten Datensätzen ( $m=\infty$ ) zu 99,7% identisch wäre mit der Parametervarianz, die sich bei Benutzung einer bestimmten Anzahl von simulierten Samples (hier:  $m=20$ ) ergibt. Dementsprechend sollte für eine optimale MI-gestützte Analyse der RE-Wert möglichst nahe an einem Wert von 1,00 liegen. Bei einem RE-Wert von unter 0,90 wäre es mithin sinnvoll, die Anzahl der Simulationen ( $m$ ) deutlich zu erhöhen.

### **Berechnung von kombinierten standardisierten Regressionskoeffizienten**

Wie aufgrund der leeren Spalte in Tabelle 9a zu erahnen ist, liefert SPSS in der von uns benutzten Version 24 für das kombinierte Ergebnis von MI-gestützten Regressionsanalysen keine standardisierten Regressionskoeffizienten (Beta bzw.  $b^*$ ). Diese Koeffizienten müssen in einem zusätzlichen, separat durchzuführenden SPSS-Analyseschritt ermittelt werden.

Zur Berechnung der standardisierten, MI-erzeugten und dann zusammengeführten Regressionskoeffizienten benutzen wir einen einfachen „Trick“: Wenn in einer Regressionsanalyse die Modellvariablen vor Durchführung der Schätzung standardisiert bzw. z-transformiert werden, sind die im Output als „nicht standardisiert“ bezeichneten Regressionskoeffizienten identisch mit den hier gesuchten standardisierten Koeffizienten. Somit müssen wir die oben berichteten Anweisungen zur Durchführung einer MI-gestützten Regressionsschätzung nur um einen dritten Block erweitern, in dem zunächst die Rohwerte der Modellvariablen in Z-Werte transformiert werden (zu erkennen an dem Buchstaben „Z“, der von SPSS an den Anfang eines jeden Variablennames gesetzt wird). Danach müssen wir dann mit den z-transformierten Variablen wiederum alle 20 separaten Regressionsschätzungen plus eine kombinierte Regressionsschätzung durchführen lassen. Dies sei hier beispielhaft verdeutlicht:

Zunächst müssen alle SPSS-Anweisungen für eine MI-gestützte Regressionsanalyse mit unseren Antisemitismusdaten ausgeführt werden. Die dafür einzusetzende SPSS-Syntax 11.1 wird auch in der folgenden SPSS-Syntax 12.1 wieder benutzt, aber dann um einen angehängten, dritten Block erweitert, mit dem die standardisierten kombinierten Koeffizientenschätzwerte ermittelt werden können. Diese sind dann allerdings, wie oben begründet, im SPSS-Output in der Spalte „Nicht standardisierte Koeffizienten“ zu finden.

### SPSS-Syntax 12.1

```

DATASET DECLARE AS_imputiert WINDOW=HIDDEN.
MULTIPLE IMPUTATION AS_Index Alter Bildung Linksrechts Geschl
/IMPUTE METHOD=FCS NIMPUTATIONS=20
/IMPUTATIONSUMMARIES MODELS
/OUTFILE IMPUTATIONS=AS_imputiert.

DATASET ACTIVATE AS_imputiert WINDOW=ASIS.
REGRESSION
/STATISTICS COEFF OUTS CI(95) R ANOVA
/DEPENDENT AS_Index
/METHOD=ENTER Alter Bildung.

DESCRIPTIVES AS_Index Alter Bildung
/SAVE.
REGRESSION
/DEPENDENT ZAS_Index
/METHOD=ENTER ZAlter ZBildung.

```

Mit SPSS-Syntax 12.1 werden die MI und zugleich auch die Regression in einem einzigen SPSS-Programmschritt durchgeführt. Falls MI und Regression in zwei, zeitlich versetzten Arbeitsschritten durchgeführt werden sollen, oder wenn der Datensatz mit den imputierten Werten in ein anderes EDV-Statistikprogramm eingelesen werden soll, muss der mittels SPSS-MI vervollständigte Datensatz ausgespeichert werden (SPSS-Syntax 12.2). Um sodann mit dem ausgespeicherten Datensatz in SPSS eine Regression durchzuführen, muss die folgende SPSS-Syntax 12.3 benutzt werden.

### SPSS-Syntax 12.2

```

DATASET DECLARE AS_imputiert WINDOW=HIDDEN.
MULTIPLE IMPUTATION AS_Index Alter Bildung Linksrechts Geschl
/IMPUTE METHOD=FCS NIMPUTATIONS=20
/IMPUTATIONSUMMARIES MODELS
/OUTFILE IMPUTATIONS='C:\Users\SPSS\AS_imputiert.sav'.

```

### SPSS-Syntax 12.3

```

GET FILE 'C:\Users\SPSS\AS_imputiert.sav'.
SORT CASES BY Imputation_.
SPLIT FILE BY Imputation_.
REGRESSION
/STATISTICS COEFF OUTS CI(95) R ANOVA
/DEPENDENT AS_Index
/METHOD=ENTER Alter Bildung.

DESCRIPTIVES AS_Index Alter Bildung
/SAVE.
REGRESSION
/DEPENDENT ZAS_Index
/METHOD=ENTER ZAlter ZBildung.

```

Die in SPSS-Syntax 12.1 und 12.3 gelisteten SPSS-Anweisungen erzeugen eine sehr umfangreiche Ausgabe mit vielen Tabellen. Erst am Ende der allerletzten Tabelle (in deren letzter Zeile) werden die kombinierten standardisierten Regressionskoeffizienten in der Spalte mit der „falschen“ Überschrift „Nicht standardisierte Koeffizienten“ ausgegeben.<sup>35</sup> Die in unserem Beispiel dort zu findenden kombinierten standardisierten Koeffizienten betragen:

$$\text{Alter, } b_1^* = -0,021$$

$$\text{Bildung, } b_2^* = -0,315$$

Zum Vergleich seien hier auch noch einmal die kombinierten unstandardisierten Koeffizientenschätzwerte aus Tabelle 9a angeführt:

$$\text{Alter, } b_1 = -0,001$$

$$\text{Bildung, } b_2 = -0,316$$

Dass in unserem Anwendungsbeispiel die Werte für den kombinierten standardisierten und für den kombinierten unstandardisierten Bildungseffekt fast identisch sind, ergibt sich im Wesentlichen daraus, dass die Standardabweichungen der abhängigen Variablen „AS\_Index“ und des Prädiktors „Bildung“ fast gleich groß sind (vgl. dazu Urban/Mayerl 2011: 96ff).

### **Berechnung von kombiniertem R<sup>2</sup>-Wert und kombiniertem F-Wert**

Die MI-Prozedur von SPSS liefert keinen kombinierten Zahlenwert für den Determinationskoeffizienten (R<sup>2</sup>) und auch nicht für den F-Wert zum Test des Gesamtmodells.<sup>36</sup> Zur Ermittlung dieser beiden Maßzahlen in kombinierter Form gibt es auch noch keine Algorithmen, die in der Sozialforschung durchgängig akzeptiert werden.<sup>37</sup> Für die Forschungspraxis schlagen wir deshalb ein ausschließlich pragmatisch ausgerichtetes Vorgehen zur Ermittlung der kombinierten R<sup>2</sup>- und F-Werte vor. Die beiden kombinierten Werte können als (arithmetische) Mittelwerte von allen m berichteten R<sup>2</sup>- und F-Werten berechnet werden. Im Folgenden sei das bezüglich des kombinierten R<sup>2</sup>-Werts verdeutlicht:

Unter den vielen Ergebnistabellen, die SPSS bei Durchführung einer MI-gestützten Regressionsanalyse ausgibt, befindet sich auch die folgende Tabelle 10. Diese Tabelle informiert zu jeder Teilregression über verschiedene Varianten des R-Wertes. Wenn für eine bestimmte R-Variante über alle Teilregressionen ein arithmetischer Mittelwert berechnet wird, so kann dieser als Proxy für den kombinierten R-Wert benutzt werden. Z.B. hat der arithmetische Mittelwert von allen zwanzig R<sup>2</sup>-Teilwerten aus Tabelle 10 einen „per Hand“ berechneten Wert von 0,098. Dieser Wert kann (in pragmatischer

<sup>35</sup> Die Formatierung der Ausgabe entspricht dem Format der oben abgedruckten Tabellen 9a und 9b. Diese Tabellen werden deshalb hier nicht noch einmal abgedruckt.

<sup>36</sup> Zu R<sup>2</sup> und F-Wert vgl. Urban/Mayerl 2011: Kap. 2.2.3 und 3.3.2.

<sup>37</sup> Zur Berechnung eines kombinierten F-Werts gibt es allerdings einen Vorschlag von J.C. Schafer (ders. 1997: 113f).



Hinsicht) als kombiniertes  $R^2$  interpretiert werden. In gleicher Weise wäre auch ein kombinierter F-Wert zu berechnen.

Tabelle 10: Die „Modellübersicht“ aus der SPSS-Ausgabe zur MI-gestützten Regressionsschätzung des Antisemitismusmodells

Imputations-Nr.	R	$R^2$	Angepasstes $R^2$	Standardfehler der Schätzung
Originaldaten	,313	,098	,074	1,18874
0				
1	,255	,065	,044	1,25484
2	,313	,098	,078	1,17147
3	,275	,076	,055	1,21430
4	,342	,117	,097	1,22941
5	,286	,082	,061	1,18394
6	,274	,075	,054	1,27353
7	,318	,101	,081	1,15245
8	,322	,104	,084	1,16050
9	,344	,118	,099	1,20196
10	,379	,144	,125	1,21709
11	,342	,117	,097	1,16011
12	,328	,108	,088	1,21078
13	,301	,090	,070	1,16384
14	,324	,105	,085	1,15882
15	,319	,101	,081	1,21601
16	,295	,087	,067	1,25909
17	,296	,088	,067	1,23579
18	,302	,091	,071	1,18252
19	,294	,086	,066	1,15912
20	,325	,106	,086	1,16652

In der folgenden Tabelle 11 werden die Ergebnisse der drei Regressionsschätzungen, die wir zur Analyse unseres Antisemitismusmodells durchgeführt haben, und für die jeweils ein anderes Verfahren zur Behandlung der fehlenden Werte beim Antisemitismusindex benutzt wurde, noch einmal gegenübergestellt. Wie die Tabelle ausweist, wurde für die Schätzung von Modell 1 ein listenweiser Ausschluss aller Fälle mit fehlenden Werten vorgenommen. Für die Schätzung von Modell 2 wurden die fehlenden Werte durch eine stochastische Regressionsimputation substituiert. Und zur Schätzung von Modell 3 wurde das Verfahren der Multiplen Imputation eingesetzt.

Die in den drei Modellen geschätzten Parameter unterscheiden sich nicht wesentlich voneinander und würden bestimmt nicht zu unterschiedlichen substanziellen Interpretationen hinsichtlich des spezifizierten Antisemitismusmodells führen. Auffällig ist in Modell 3, dass dort zwei der drei Standardfehler (mit Werten von 0,625 und 0,116) größer sind als in Modell 1 und Modell 2. Dies ist eine gewünschte Folge der Konstruktionslogik des MI-Verfahrens, welche, wie oben ausgeführt, die erhöhte Unsicherheit, die bei Modellschätzungen mit imputierten Werten auftritt, bei der Berechnung von Standardfehlern und von Signifikanzniveaus (die sich in Modell 3 teilweise sehr deutlich von denen in Modell 2 unterscheiden) berücksichtigt.

Zu beachten ist, dass diese Ergebnisse nur illustrative Bedeutung haben. Mit einer pro Modell nur einmal durchgeführten Schätzung können keine systematischen Ergebnisse gewonnen werden, die etwas über die Qualität des jeweils eingesetzten Verfahrens beim Umgang mit fehlenden Werten aussagen. Insbesondere sind auf diese Weise keine Aussagen darüber zu gewinnen, ob die vorgestellten Verfahren prinzipiell unverzerrte, effiziente und konsistente Schätzwerte (vgl. dazu Urban/Mayerl 2011: Kap. 3.1.1) in Regressionsanalysen mit fehlenden Werten ermöglichen. Dazu wären Methodenstudien mit einer sehr großen Anzahl von Samples unter systematisch variierten Bedingungen (hinsichtlich: Missingquote, Fallzahl, m, Verteilungsform usw.) notwendig. Dementsprechend basieren in diesem Lehrbuch auch alle Aussagen über die grundsätzlichen Eigenschaften von Verfahren zum Umgang mit fehlenden Werten nicht auf den Ergebnissen, die wir mit unserem Beispieldatensatz erzielt haben, sondern auf den Ergebnissen von systematischen Studien, die oftmals auch in den dazugehörigen Fußnoten benannt werden. Ein gelungenes Beispiel für eine solche systematische Studie zur Überprüfung der Qualität des MI-Verfahrens liefern Graham/ Schafer 1999.

**Tabelle 11:** Ergebnisse der Regressionsschätzung (Modell 0) ohne Missings (nur „complete cases“) und von drei Schätzungen mit unterschiedlichen Verfahren zur Behandlung fehlender Werte (Modelle 1 - 3) (Antisemitismusmodell)

	Modell 0	Modell 1	Modell 2	Modell 3
	„complete cases“	„listwise deletion“	stoch. Regr.-imputation	Multiple Imputation
Brutto-N	92	92	92	92
Missingquote	---	16,3%	vor Imputation: 16,3%; nach Imputation: 0,0%	vor Imputation: 16,3%; nach Imputation: 0,0%
Netto-N	92	77	92	92
a				
Konst.	3,855	4,033	4,309	4,069
b				
Alter	0,003	-0,001	-0,007	-0,001
Bildung	-0,329	-0,315	-0,302	-0,315
SE				
Konst.	0,610	0,618	0,604	0,625
Alter	0,008	0,008	0,008	0,008
Bildung	0,109	0,114	0,108	0,116
b*				
Alter	0,045	-0,012	-0,101	-0,021
Bildung	-0,322	-0,316	-0,304	-0,315
Sign. (p)				
Konst.	0,000	0,000	0,000	0,000
Alter	0,670	0,918	0,357	0,828
Bildung	0,003	0,007	0,006	0,005
R-Quadrat	0,116	0,098	0,081	0,098

### Runden von intervallskalierten Daten

Wenn fehlende Werte bei intervallskalierten Variablen (wie z.B. bei der Variablen „Bildung“ mit Werten von 1 bis 5) mittels MI-Verfahren substituiert werden, sind die imputierten Variablenwerte fast immer Dezimalzahlen. So wird dann z.B. nicht ein Wert von 5,00, sondern ein Wert von 4,65 imputiert. Wenn dies nicht erwünscht ist (z.B. wenn sich die Regressionsanalyse auf stetige Variablen mit natürlichen Werten beziehen soll), kann SPSS gezwungen werden, die zu imputierenden Zahlen auf ganze Zahlenwerte zu runden. Dazu muss die in SPSS-Syntax 11.1 (sowie in Syntax 11.2 und Syntax 12.1/12.2) benutzte SPSS-Anweisung „MULTIPLE IMPUTATION ...“ um einen Unterbefehl erweitert werden, der im Falle der Variablen „Bildung“ wie folgt aussehen würde:

#### SPSS-Syntax 13

```
/CONSTRAINTS Bildung (MIN=1 MAX=5 RND=1)
```

Generell gilt: Mit dem Runden von zu imputierenden Werten geht Information und damit die Schätzgenauigkeit verloren. Deshalb sollte nur dann gerundet werden, wenn eine Variable allein ganzzahlige (natürliche) Werte aufweisen soll. Wann immer vertretbar, sollten die im MI-Verfahren geschätzten Werte nicht gerundet werden.<sup>38</sup> In unserem Antisemitismusdatensatz haben wir z.B. die zu imputierenden Werte bei der abhängigen Variablen „AS\_Index“ nicht gerundet. Auch wenn durch die Kombination von ganzen und gebrochenen Zahlenwerten im Wertebereich einer Variablen ein „unschönes“ Verteilungsbild entsteht, in dem neue Dezimalzahlen auftauchen, die nur wenige Fälle repräsentieren, sollte von einer Rundung eher abgesehen werden.

Wenn fehlende Werte von kategorialen Variablen im MI-Verfahren geschätzt werden (s.u.), so stellt sich die Rundungsfrage überhaupt nicht, denn die fehlenden Werte werden von SPSS automatisch als ganze Zahlen imputiert.

### Multiple Imputation mit kategorialen Variablen

Das in SPSS implementierte MI-Verfahren kann umstandslos auch fehlende Werte von kategorialen Variablen substituieren. Zur Gruppe der kategorialen Variablen gehören:

- a) binomiale bzw. binäre Variablen, wie z.B. die Variable „Beschäftigungsstatus“ mit den Ausprägungen „beschäftigt/nicht beschäftigt“,

---

<sup>38</sup> Frühere Auffassungen zur Notwendigkeit des Rundens im MI-Verfahren (vgl. Schafer/Olsen 1998: 550) werden durch die neuere Methodenforschung korrigiert. Vgl. dazu: Enders 2010: 261f; Graham 2009: 561; Von Hippel 2009.

- b) multinomiale kategoriale Variablen, wie z.B. die Variable „Parteipräferenz“ mit den Ausprägungen „CDU/ SPD/ Grüne/ andere“,
- c) geordnet kategoriale bzw. ordinale Variablen, wie z.B. die Variable „Hochschulabschluss“ mit den Ausprägungen „keiner/ BA/ MA/ Promotion“.

Damit in SPSS das MI-Verfahren bei kategorialen Variablen reibungslos funktioniert, muss die betreffende kategoriale Variable unbedingt als Variable mit nominalem oder ordinalem Messniveau definiert sein. Ist das nicht der Fall und wird eine kategoriale Variable irrtümlich als metrische Variable im SPSS-Datensatz definiert, so kann das nachträglich u.a. durch folgende SPSS-Anweisung korrigiert werden (hier beispielhaft für die Variable „Praeferenz“ mit nominalem Messniveau veranschaulicht):

#### SPSS-Syntax 14

```
VARIABLE LEVEL Praeferenz(NOMINAL).  
EXECUTE.
```

Wenn fehlende Werte von kategorialen Variablen im MI-Verfahren geschätzt werden, so müssen die zu imputierenden Werte nicht, wie oben erläutert, zusätzlich gerundet werden. Die fehlenden Werte werden von SPSS als ganze Zahlen imputiert. Voraussetzung dafür ist natürlich, dass das Messniveau der betreffenden Variablen als „nominal“ oder „ordinal“ definiert wurde (s.o.).

Kategoriale Variablen, deren fehlende Werte mittels des SPSS-MI-Verfahrens in *m* Datensätzen substituiert wurden, können in SPSS in unterschiedlichsten statistischen Modellierungen als abhängige Variable analysiert werden. Zu diesen Verfahren gehören u.a.:

- die „binäre logistische Regression“,
- die „multinomiale logistische Regression“,
- die „ordinale Regression“.

Für diese Analyseverfahren kann SPSS die im MI-Verfahren angeforderte Anzahl von „*m*“ Einzelschätzungen und eine zusammengeführte bzw. kombinierte Gesamtschätzung liefern.

Fassen wir zusammen: Das MI-Verfahren hat Vorteile (+) und Nachteile (-):

- (+) Mittels MI-Verfahren (FCS-Methode) können sowohl bei metrischen Variablen (mindestens intervallskaliert) als auch bei kategorialen Variablen (bi- oder multinomial, geordnet oder ungeordnet)<sup>39</sup> die fehlenden Werte durch neu geschätzte Werte substituiert werden.
- (+) Die MI kann bei fehlenden Werten mit MAR-Eigenschaft unverzerrte Schätzwerte von Parametern und Standardfehlern liefern. Und auch unter MCAR-Bedingungen kann die MI sehr gute Ergebnisse liefern.
- (+) Die MI kann auch bei hohen Missingquoten (z.B. bei Variablen mit 40% fehlenden Werten), bei kleinen Fallzahlen (z.B. bei einem Netto-N von 50 Fällen) und bei großen Regressionsmodellen (z.B. mit  $k=18$  unabhängigen Variablen) eingesetzt werden.<sup>40</sup>
- (+) Für die MI gilt, dass sie in aller Regel bei deutlich nicht-normalverteilten Variablen gute Ergebnisse liefert. Anzumerken ist, dass dies auch für Verfahren (andere als die beschriebene FCS-Methode) gilt, die eine multivariate Normalverteilung<sup>41</sup> aller beteiligten Variablen explizit voraussetzen<sup>42</sup>.
- (+) Im MI-Verfahren werden die imputierten Werte nicht als tatsächliche, sondern als methodisch erzeugte Daten betrachtet, deren Verwendung mit erheblichen Unsicherheiten verbunden ist. Deshalb wird die Varianz der Parameterschätzungen unter Berücksichtigung von m-Schätzungen ermittelt. Bei der Berechnung von Standardfehlern und Irrtumswahrscheinlichkeiten wird auch die Unsicherheit der Imputation berücksichtigt.
- (+) Zur Schätzung der zu imputierenden Werte können im MI-Verfahren in der Imputationsphase (Schritt 1 des Verfahrens, s.o.) auch multinomial skalierte Prädiktoren (kategorial geordnet oder kategorial ungeordnet) benutzt werden.<sup>43</sup> Auch können binomiale/binäre Prädiktoren und Hilfsvariablen (Dummies) eingesetzt werden.

<sup>39</sup> Andere Softwareprogramme arbeiten ggf. mit anderen Methoden und setzen dann eine multiple Normalverteilung aller Variablen voraus. Obwohl diese Methoden (formallogisch betrachtet) eigentlich nur bei metrischen Variablen einsetzbar sind, können sie zum Teil auch adäquate Ergebnisse bei kategorialen Variablen liefern.

<sup>40</sup> Vgl. Graham 2009: 560.

<sup>41</sup> Um kategoriale Variablen zu berücksichtigen müssten diese dort in Form von  $k-1$  Dummies in das Modell aufgenommen werden.

<sup>42</sup> Vgl. dazu Cole 2008: 223; Enders 2010: 259f; Graham 2009: 560; Graham/Hofer 2000: 218; Schafer/Graham 2002: 167-170; Schafer/Olsen 1998: 550. Bei sehr stark schief verteilten Variablen sollten diese vor Durchführung eines MI-Verfahrens transformiert werden (z.B. mittels einer Box-Cox-Transformation, vgl. dazu Osborne 2013: 175-188).

<sup>43</sup> Nehmen wir zur Verdeutlichung an, die Modellvariable „ $X_1$ “ sei kategorial und die Modellvariablen „ $X_2$ “ und „ $X_3$ “ seien metrisch skaliert. Dann werden als erstes die Missings in  $X_1$  unter Verwendung eines logistischen Schätzmodells ersetzt. Im Anschluss daran werden die Missings in  $X_2$  mit  $X_3$  und  $X_{1+i}$  ersetzt (vgl. Anhang). Da aber  $X_1$  kategorial ist, muss SPSS diese als Dummies verwenden. Die dementsprechende Rekodierung macht das Programm automatisch. Somit können also auch kategoriale Variablen als Prädiktoren in das MI-Verfahren einbezogen werden.

- (–) Die fehlenden Werte müssen formallogisch betrachtet mindestens MAR-Eigenschaft aufweisen. Unter MNAR-Bedingungen gibt es keine Gewissheit, dass das MI-Verfahren zuverlässige und robuste Schätzergebnisse liefert.<sup>44</sup>
- (–) Die Schätzung von standardisierten Regressionskoeffizienten muss in SPSS noch in einem zusätzlichen Analysedurchlauf separat durchgeführt werden. Und zur Berechnung von kombinierten bzw. zusammengefassten Determinationskoeffizienten ( $R^2$ ) und F-Werten stehen softwareunabhängig noch keine angemessenen Prozeduren, sondern nur pragmatische Verfahrensvorschläge zur Verfügung.

## 4 Imputation und Design-Gewichtung in SPSS

Wenn Datensätze nachträglich gewichtet werden müssen, dann ist besondere Sorgfalt beim Einsatz von Imputationsverfahren aufzuwenden. Warum? Im Nachfolgenden soll der Gebrauch von Gewichtungsvariablen und die gleichzeitige Imputation von Variablenwerten zur Substitution von Missings bei Verwendung des Statistikprogrammpakets SPSS erläutert werden.

Oftmals sind Survey-Datensätze hinsichtlich der darin enthaltenen Fallzahlen von Personen, die bestimmten sozialen Gruppen angehören, disproportional geschichtet. Es gibt dann in einer Gruppe viel mehr Befragte als eigentlich darin sein dürften und die relative Sample-Größe der betreffenden Gruppe entspricht nicht der relativen Größe dieser Gruppe in der Population. So wird z.B. im ALLBUS(2014)-Datensatz, der alle zwei Jahre als nationaler allgemeiner Survey der gesamtdeutschen Bevölkerung erhoben wird, ein Anteil von in Ostdeutschland lebenden Personen befragt (ab 18 Jahre, in Privathaushalten wohnend), der ca. 32% aller insgesamt für den ALLBUS befragten Personen entspricht. Tatsächlich leben aber nur ca. 18% aller Deutschen (mit den zuvor genannten Merkmalen) in Ostdeutschland. Die Gruppe der für den ALLBUS befragten, ostdeutschen Personen ist also wesentlich umfangreicher als sie eigentlich sein dürfte, wenn mit den ALLBUS-Daten repräsentative Aussagen über alle Personen, die in Ost- und Westdeutschland (also in Gesamtdeutschland) leben, berechnet werden sollen.<sup>45</sup>

---

<sup>44</sup> Vgl. Osborne 2013: 125-129; Scheffer 2002: 160. In den Simulationsstudien von Schafer/Graham (dies. 2002: 169) wurde z.B. unter MNAR-Bedingungen ein tatsächlicher Regressionskoeffizient von 0,60 in 1000 MI-Samples nur mit einem durchschnittlichen Wert von 0,21 geschätzt. Und die Abdeckung ("Coverage") lag dabei nur bei 39% (sie sollte mindestens 95% betragen). Als "Coverage" wird der prozentuale Anteil aller Konfidenzintervalle in allen Regressions-schätzungen bezeichnet, die den wahren Regressionskoeffizienten, der in den MI-Verfahren geschätzt werden musste, enthalten (Schafer/Graham 2002: 169).

<sup>45</sup> Für weitere Informationen zur Design-Gewichtung im ALLBUS 2014 vgl. Schulz 2015: iii-iv.

Eine solche Überrepräsentation bestimmter Befragtengruppen erfolgt in Umfragen immer dann, wenn mit den Befragungsdaten spezielle Analysen durchzuführen sind, die nur für die überrepräsentierte Gruppe gelten sollen, und für die ohne Oversampling zu wenige Beobachtungsfälle zur Verfügung ständen. Denn wäre die Anzahl befragter Ostdeutscher entsprechend ihres relativen Anteils an der deutschen Bevölkerung ermittelt worden, hätten von dieser Gruppe nur 625 Personen in den Survey aufgenommen werden dürfen. Es ist aber unmittelbar einsichtig, dass Datenanalysen mit den Angaben von 1109 Personen anspruchsvollere und auch qualitativ bessere Auswertungen ermöglichen als Analysen mit den Angaben von nur 625 Personen (zumindest gilt dies im Prinzip). Deshalb werden beim Sampling für Surveystudien oftmals bestimmte Bevölkerungsgruppen überrepräsentativ erfasst.

Wenn jedoch die Daten von disproportional geschichteten Stichproben insgesamt ausgewertet werden sollen, sodass die Analyseergebnisse nicht nur für die überrepräsentierte Gruppe, sondern für alle Befragten gültig sind, muss das Oversampling wieder rückgängig gemacht werden. Dazu wird in der Forschung eine Design-Gewichtung eingesetzt. Dadurch wird z.B. im hier erörterten ALLBUS 2014 den ostdeutschen Befragten ein geringeres Gewicht und den westdeutschen Befragten ein höheres Gewicht zugewiesen. Dies wird dadurch erreicht, dass in der Datenanalyse eine Gewichtungsvariable (im ALLBUS 2014: v870) eingesetzt wird, die nur zwei Ausprägungen hat (0,5639113 und 1,2047504). Die Daten eines jeden ostdeutschen Befragten werden sodann mit einem Gewicht von 0,5639113 und die Daten eines jeden westdeutschen Befragten mit einem Gewicht von 1,2047504 versehen. Auf diese Weise verändert sich die disproportionale Schichtung des ALLBUS-Surveys von 32:68 Prozent (ostdeutsche zu westdeutsche Befragte) wieder zu der richtigen proportionalen Schichtung von 18:72 Prozent, ohne dass dadurch die Gesamtzahl aller Befragten (3471 Personen) verändert wird.

Rein technisch betrachtet wird das Oversampling im ALLBUS 2014 durch die folgende SPSS-Anweisung 15 mit der Gewichtungsvariablen „v870“ (aus dem ALLBUS 2014) rückgängig gemacht:

#### SPSS-Syntax 15

WEIGHT BY v870.

Allerdings, und das ist wichtig (!), muss in SPSS zuerst das Imputationsverfahren mit ungewichteten Daten (egal ob mittels sRI- oder MI-Verfahren) und dann erst die Samplegewichtung für die weitere statistische Datenanalyse ausgeführt werden.

Wenn diese Reihenfolge nicht eingehalten wird, werden z.B. bei eingeschalteter Gewichtung und gleichzeitiger Anwendung der stochastischen Regressionsimputation alle Fälle aus dem jeweiligen SPSS-Datensatz gelöscht, die ein Samplegewicht von unter 1,0 aufweisen. Das hieße, dass der imputierte Datensatz nur noch diejenigen Fälle enthielte, denen vor der Imputation ein Gewicht von mindestens 1,0 zugeordnet wurde. Denn SPSS löscht die anderen Fälle, weil es bei eingeschalteter Gewichtung und gleichzeitiger Imputation nicht zur nächstgelegenen Zahl auf- oder abrundet, sondern grundsätzlich immer abrundet. Bei einem definierten Gewicht von z.B. 1,9 wird SPSS dieses Gewicht auf einen Wert von 1,0 abrunden. Entsprechend verhält sich SPSS bei einem definierten Gewicht von 0,9. Auch dieses Gewicht wird auf einen Wert von 0,0 gerundet. Und damit werden die davon betroffenen Fälle aus der weiteren Analyse ausgeschlossen.

Für die oben erläuterte ALLBUS-Gewichtung mit den beiden Gewichten von 0,5639113 und 1,2047504 bedeutete dies z.B., dass damit alle ostdeutschen Fälle aus dem imputierten Datensatz gelöscht würden. Denn wenn Fälle mit einem Wert von 0,5639113 gewichtet werden, wird die Gewichtung durch SPSS automatisch auf 0,0 abgerundet. Es bliebe dementsprechend in unserem Beispiel nur ein imputierter ALLBUS-Datensatz übrig, der allein die Daten von Personen aus der westdeutschen Bevölkerung enthielte.

Eine ähnliche Problematik ist auch bei der Kombination von Gewichtung und Multipler Imputation in SPSS zu beobachten. Dabei werden die Gewichte stets zur nächstgelegenen Integerzahl gerundet, was z.B. dazu führt, dass Fälle mit einem Gewicht von 0,3 nicht imputiert und Fälle mit einem Gewicht von 0,7 dagegen doch imputiert werden. Dementsprechend werden auch Fälle mit einem Gewicht von 0,0 nicht imputiert, während Fälle mit einem Gewicht von 0,5 sehr wohl imputiert werden. Auch werden keine Werte imputiert, wenn die Fälle ein negatives Gewicht aufweisen.<sup>46</sup>

Grundsätzlich sollte also bei Verwendung von SPSS zur Imputation fehlender Werte eine evtl. notwendige Gewichtung erst nach der Imputation durchgeführt werden.

In welcher Weise die nachträgliche Gewichtung bei Durchführung eines MI-Verfahrens erfolgen kann, zeigt die folgende SPSS-Syntax 16. Dabei beziehen wir uns auf die im SISS-Kapitel 3.3 abge-

---

<sup>46</sup> Mit negativen Gewichten lassen sich in SPSS generell keine statistischen Prozeduren (bspw. auch keine Häufigkeitsauszählungen) durchführen.



druckten Anweisungen zur Durchführung eines MI-Verfahrens (SPSS-Syntax 12.1). Die Syntax besteht dort aus drei Blöcken, deren erste Zeilen im Folgenden noch einmal angezeigt werden (SPSS-Syntax 16).

### SPSS-Syntax 16

```
DATASET DECLARE AS_imputiert WINDOW=HIDDEN.  
MULTIPLE IMPUTATION AS_Index Alter Bildung Linksrechts Geschl  
usw.  
  
DATASET ACTIVATE AS_imputiert WINDOW=ASIS.  
REGRESSION  
usw.  
  
DESCRIPTIVES AS_Index Alter Bildung  
/SAVE.  
usw.
```

Für eine nachträgliche Gewichtung im MI-Verfahren sollte in SPSS vor Block 2 und (zur Sicherheit) vor Block 3 die nachfolgende Anweisung (SPSS-Syntax 17) eingeschoben werden. Dabei bezeichnet „AS\_imputiert“ (der Name wurde in Block 1 vergeben) den Datensatz, der die zuvor imputierten Werte enthält. Und die Bezeichnung „g1“ benennt den Namen für die jeweils benutzte Gewichtungsvariable.

### SPSS-Syntax 17

```
DATASET ACTIVATE AS_imputiert.  
WEIGHT BY g1.
```

## **Anhang: Multiple Imputation mit SPSS und der Fully Conditional Specification Methode (FCS-Methode)**

Wie in SISS-Kapitel 3.3 erläutert wurde, werden im MI-Verfahren während der Imputationsphase die fehlenden Werte einiger oder aller Modellvariablen durch geschätzte Werte ersetzt<sup>47</sup>. Die Auswahl der Methode zur Ersetzung der Missings ist dabei von deren Verteilung im Datensatz abhängig, dem sog. „Missingmuster“. Für SPSS ist wichtig, ob ein generelles bzw. rein zufälliges oder ein monotonen Missingmuster vorliegt. Bei einem generellen/ zufälligen Missingmuster (s.u.) wird von SPSS zur Schätzung der fehlenden Werte die iterative *Fully Conditional Specification* Methode (FCS) und bei einem monotonen Missingmuster (s.u.) die *Montone* Methode verwendet.

Bei einem generellen Missingmuster darf die Verteilung der Missings im Vergleich zwischen den Fällen und den Variablen keinerlei Systematik aufweisen. Die Missings sollten dann rein zufällig im Datensatz verteilt sein. Eine solche zufällige Verteilung wird beispielhaft in der folgenden Tabelle A1 (linker Teil) veranschaulicht. Dort hat der Fall 6 einen fehlenden Wert in  $X_2$ , ist aber ansonsten vollständig vorhanden. Der Fall 5 hingegen weist vier fehlende Werte auf (in  $X_2$ ,  $X_3$ ,  $X_4$  und  $X_8$ ). Somit weist die Verteilung der fehlenden Werte im Vergleich zwischen Fall 5 und 6 und zwischen den Variablen keine Systematik auf. Ein Zusammenhang zwischen den Missing-Verteilungen bei diesen beiden Fällen ist nicht zu erkennen und somit liegt hier ein generelles bzw. zufälliges Missingmuster vor.

Anders ist es beim monotonen Missingmuster. Solch ein Muster ist vor allem bei Längsschnittdaten zu beobachten. Da z.B. in Panelstudien oftmals einige im Zeitverlauf befragte Personen noch vor Abschluss der Studie ihre Studienbeteiligung abbrechen (Panelmortalität) und somit für diese Befragten ab einem bestimmten Zeitpunkt nur noch fehlende Werte zu registrieren sind, können dort leicht monotone Missingmuster entstehen. Eine solche systematisch-monotone Missingverteilung zeigt Tabelle A1 (rechter Teil). Fehlende Werte sind in Tabelle A1 mit (-) und beobachtete Werte mit (+) gekennzeichnet. Ein monotones Muster ist immer dann gegeben, wenn die Variablen nach ihrer Missinganzahl geordnet werden können und wenn (bezogen auf das Beispiel) das Folgende gilt: Wenn für einen Fall der Variablenwert von  $X_8$  beobachtet wurde, müssen auch die Werte aller vorherigen Variablen  $X_{1...7}$  beobachtet sein. Umgekehrt muss das nicht so sein. Sollte für einen Fall der Wert von  $X_8$  fehlend sein, könnten auch die Werte von  $X_{1...7}$  fehlen, müssen es aber nicht. SPSS wendet bei solchen monotonen Missingmustern die *Monotone Method* an. Diese Methode wird hier, weil sie

---

<sup>47</sup> Die nachfolgenden, die Software SPSS betreffenden Ausführungen beziehen sich auf: IBM SPSS Statistics 22 Algorithms (2013: 597 – 602).

eher ein Spezialfall von Datenverteilung ist und wir uns hier auf Querschnittsdaten konzentrieren wollen, nicht näher beschrieben<sup>48</sup>.

**Tabelle A1:** Generelles/zufälliges und monotonen Missingmuster

Generelles Missingmuster									Monotonen Missingmuster <sup>49</sup>								
	X1	X2	X3	X4	X5	X6	X7	X8	X1	X2	X3	X4	X5	X6	X7	X8	
1	+	+	-	+	+	+	-	-	+	+	+	+	+	+	+	+	
2	-	+	+	+	+	-	+	+	+	+	+	+	+	+	+	-	
3	-	+	+	+	+	-	+	+	+	+	+	+	+	+	-	-	
4	+	+	+	+	-	+	-	+	+	+	+	+	+	-	-	-	
5	+	-	-	-	+	+	+	-	+	+	+	+	-	-	-	-	
6	+	-	+	+	+	+	+	+	+	+	+	-	-	-	-	-	
7	+	+	-	+	+	+	+	-	+	+	-	-	-	-	-	-	

Zu beachten ist, dass das Missingmuster unabhängig von den Missingmechanismen ist. Das Muster zielt nur auf die Verteilung der fehlenden Werte im Datensatz ab und nicht darauf, ob diese von anderen Variablen abhängig oder unabhängig sind: unter einem zufälligen Missingmuster liegt damit nicht unbedingt MCAR vor.

Im Folgenden soll die FCS-Methode genauer beschrieben werden, die SPSS bei einem allgemeinen/zufälligen Missingmuster zur Schätzung der fehlenden Werte immer dann automatisch einsetzt, wenn es kein monotonen Muster entdecken kann<sup>50</sup>. Wir werden zuerst die FCS-Methode in ihren Grundzügen charakterisieren und dann die einzelnen Schritte der Verfahrensweise im Detail erläutern.

Zur Durchführung einer FCS-Schätzung müssen die Variablen im Datensatz entsprechend ihrer Skalenniveaus richtig spezifiziert sein. Das heißt, dass metrische, ordinale und nominale Variablen in der SPSS-Datenmaske als solche definiert sein müssen (evtl. nachträglich mit der SPSS-Syntax 14, s.o.). Das ist notwendig, weil die FCS-Methode für jede Variable ein eigenständiges Imputationsmodell anlegt, welches wiederum vom Skalenniveau der jeweiligen Variablen abhängig ist: Zur Ersetzung

<sup>48</sup> Grundsätzlich läuft die Ersetzung der Missings in einem solchen Muster wie folgt ab (hier für einen Datensatz mit X<sub>1</sub> ... X<sub>8</sub> Variablen erläutert). Zuerst werden die Missings für diejenige Variable mit den wenigsten Missing Values (hier X<sub>3</sub>) anhand der vorhandenen Information von X<sub>1</sub> und X<sub>2</sub> ersetzt. Man erhält eine imputierte Variable X<sub>3+1</sub>. Danach werden die Missings in X<sub>4</sub> ersetzt (in diesem Fall geschieht das mit den Informationen aus X<sub>1</sub>, X<sub>2</sub> und X<sub>3+1</sub>). Dies wird solange wiederholt, bis auch X<sub>8</sub> ersetzt ist. Anzumerken ist, dass für jedwede Variable ein eigenes Imputationsmodell, entsprechend des Skalenniveaus, angelegt wird (wie dies bei FCS der Fall ist).

<sup>49</sup> Eigene Darstellung in Anlehnung an van Buuren (2007: 225).

<sup>50</sup> Ab SPSS-Version 17.

von Missings in metrischen Variablen benutzt die FCS-Methode per Voreinstellung ein lineares Regressionsmodell (der Anwender hat aber auch die Möglichkeit, statt eines linearen Regressionsmodell ein sog. „*predictive mean matching* Modell“ auszuwählen). Bei nominalen und ordinalen Variablen kommt ein logistisches Regressionsmodell zum Einsatz.

Der Imputationsprozess läuft in zwei Schritten ab: Zuerst werden die fehlenden Werte im Datensatz mit Platzhaltern gefüllt. Es liegt dann ein vervollständigter, mit Platzhaltern gefüllter Datensatz vor. Danach startet die FCS-Methode mit der Ersetzung der fehlenden Werte. Dazu werden die Platzhalter einer Variablen gelöscht und alle anderen Variablen herangezogen, um die dann entstandenen (bzw. ursprünglich vorliegenden) Missings durch Imputationswerte zu ersetzen. Die FCS-Methode führt dieses Löschen und Ersetzen der Platzhalter in iterativer Weise so lange durch, bis alle Platzhalter im Datensatz mit geschätzten Werten ersetzt sind.

Dabei ist zu beachten, dass nach den ersten Iterationsschritten die ersetzten Werte noch immer sehr stark von den Platzhaltern abhängig sind (s.u.). Es müssen mehrere Iterationen durchlaufen werden, bis die Unabhängigkeit der ersetzten Werte von den Platzhaltern sichergestellt ist. Mittels Voreinstellung legt SPSS eine Anzahl von 10 Iterationen fest. Es können aber auch weitaus mehr Iterationen gewählt werden<sup>51</sup>.

In jedem Iterationsschritt löscht der FCS-Algorithmus die ersetzten Missing aus der vorausgegangenen Iteration: So werden z.B. die imputierten Werte in  $X_1$  gelöscht und mit allen anderen Variablen erneut vorhergesagt, dann werden die imputierten Werte in  $X_2$  gelöscht und mit allen anderen Variablen (auch mit der ersetzten  $X_1$ -Variable aus diesem Iterationsschritt) vorhergesagt etc. (s.u. für eine detaillierte Erklärung). Dieses Löschen und Ersetzen wird so lange wiederholt, bis die vorgegebene Anzahl von Iterationen durchlaufen ist. Bei 10 Iterationen löscht und ersetzt der FCS-Algorithmus alle ursprünglich aufgefüllten Werte insgesamt zehnmal. Hat der FCS-Algorithmus alle Iterationen (eine Iterationskette) durchlaufen, liegt *ein* vollständig imputierter Datensatz vor, der statistisch analysiert werden kann (z.B. mittels Regressionsanalyse). Dann wurde allerdings nur eine Einfach-Imputation durchgeführt (mit einer bestimmten Anzahl von Imputationsschritten). In einer Multiplen Imputation wird demgegenüber der komplette Prozess m-mal wiederholt und es werden m selbstständige Datensätze erzeugt.

---

<sup>51</sup> Die Anzahl der Iterationen muss so gewählt werden, dass sich FCS stabilisiert. Wann dies zutrifft ist von Fall zu Fall verschieden. Grundsätzlich sollten 10 Iterationen allerdings ausreichen (es sollte aber immer geprüft werden, ob dies der Fall ist).

Im Folgenden sollen nun die einzelnen Schritte der FCS-Methode detailreicher erläutert werden:<sup>52</sup> Die Platzhalter für metrische Variablen werden aus einer Normalverteilung gezogen, die auf den beobachteten Teilen einer Variablen  $X_i$  beruht (genannt:  $X_{i,obs}$ )<sup>53</sup>. Die fehlenden Werte in  $X_i$  (genannt:  $X_{i,mis}$ ) werden dabei ignoriert. Aus den beobachteten Werten der Variablen ( $X_{i,obs}$ ) werden deren Mittelwert und Standardabweichung berechnet, um damit eine Normalverteilung für diese Variable zu spezifizieren. Die Platzhalter werden dann zufällig aus dieser Normalverteilung gezogen.

Bei kategorialen Variablen wird ebenfalls der beobachtete Teil der Variablen ( $X_{j,obs}$ ) zur Schätzung der Platzhalter herangezogen (auch hierbei werden die Missings ignoriert). Dabei interessieren die prozentual beobachteten Anteilswerte in den jeweiligen Kategorien der Variablen. Die Platzhalter werden dann auf Grundlage dieser Anteilswerte aus einer Multinomialverteilung gezogen. Die Wahrscheinlichkeit des Platzhalters, aus einer der Kategorien gezogen zu werden, entspricht dabei dem empirischen Anteilswert des beobachteten Teils dieser Variablen.

Da SPSS für die m-Datensätze einer Multiplen Imputation jeweils eine eigene Iterationskette anlegt, unterscheiden sich auch die Platzhalter von Kette zu Kette und damit auch die zur Verfügung stehenden Informationen zu Ersetzung der fehlenden Werte (s.u.).

Nachdem die Platzhalter bestimmt wurden, beginnt die Schätzung der zu imputierenden Werte nach der FCS-Methode. Folgendes Beispiel (fiktiv) soll zur Veranschaulichung des FCS-Prozesses dienen: Angenommen wird, dass für jede der acht Variablen ( $X_1 \dots X_8$ ) eines Datensatzes sowohl beobachtete Werte (+) als auch fehlende Werte (-) vorhanden sind. Zusätzlich sollen sich die Missings nach einem generellen/zufälligen Missingmuster verteilen (vgl. Tabelle A2).

---

<sup>52</sup> Wie im Folgenden erläutert wird, beruht die Ersetzung der fehlenden Werte nach der FCS-Methode auf einem zweistufigen Auswahlprozess. Dieser beruht auf Prinzipien bayesianischer Statistik: es werden zuerst aus einer posterior Verteilung bestimmte Parameter gezogen, bevor aus einer weiteren posterior Verteilung die schlussendlichen Werte für die jeweiligen Missings hervorgehen. Das hier vorgestellte Vorgehen der FCS-Methode wird ohne Bayes-Modell sehr vereinfacht dargestellt. Verstehbare Einführungen in die Bayesschen Grundlagen und Techniken der MI gibt es in: Allison 2002: 27-88; Enders 2010: 187-286.

<sup>53</sup> Der beobachtete Teil einer Variablen wird mit dem Suffix *obs* der unbeobachtete Teil einer Variablen mit dem Suffix *mis* versehen.

**Tabelle A2:** Struktur eines Datensatzes mit beobachteten (+) und fehlenden (-) Werten von acht Variablen

	X1	X2	X3	X4	X5	X6	X7	X8
1	+	+	-	+	+	+	-	-
...	-	+	+	+	+	-	+	+
n	-	+	+	+	+	-	+	+

Im ersten Schritt ersetzt der FCS-Algorithmus, wie oben erläutert, alle Missings (-) mit einem zufällig gezogenen Platzhalter (\*). Dies wird in Tabelle A3 dargestellt.

**Tabelle A3:** Datensatz mit Platzhaltern (\*) und beobachteten Variablenwerten (+)

	X1	X2	X3	X4	X5	X6	X7	X8
1	+	+	*	+	+	+	*	*
...	*	+	+	+	+	*	+	+
n	*	+	+	+	+	*	+	+

Danach löscht der FCS-Algorithmus diese Platzhalter in der ersten Variablen X<sub>1</sub>. Die Werte der Variablen X<sub>1</sub> lassen sich sodann in beobachtete Werte (X<sub>1,obs</sub>) und unbeobachtete Werte (X<sub>1,mis</sub>) aufteilen. Die Tabelle A4 verdeutlicht diese Aufteilung bei der Variablen X<sub>1</sub>.

**Tabelle A4:** Datensatz mit Platzhaltern (\*) und einer Aufteilung der X<sub>1</sub>-Werte in beobachtete Werte (+) und unbeobachtete Werte (-)

	X1	X2	X3	X4	X5	X6	X7	X8
1	+	+	*	+	+	+	*	*
...	-	+	+	+	+	*	+	+
n	-	+	+	+	+	*	+	+

Anschließend benutzt der FCS-Algorithmus, je nachdem in welcher Weise die Variable X<sub>1</sub> skaliert ist, entweder ein lineares (bzw. ein PMM-Modell)<sup>54</sup> oder ein logistisches Regressionsmodell, um die unbeobachteten Werte von X<sub>1</sub> zu schätzen. Dies wird mit der folgenden Gleichung in vereinfachter Form (ohne Gewichtungen und ohne Zufallskomponente) veranschaulicht:<sup>55</sup>

$$X_{1,mis} = X_{1,obs} + X_2 + X_3 \dots X_8$$

<sup>54</sup> Auf das Predictive Mean Matching-Verfahren (PMM-Modell) wird im vorliegenden Text nicht eingegangen. Eine kurze Charakterisierung (und Kritik) findet sich in Allison 2015.

<sup>55</sup> Bei der linearen und der logistischen Regression werden auch Zufallskomponenten in die Schätzung aufgenommen (hier nicht dargestellt).

Die unbeobachteten Werte in  $X_1$  werden demnach mit Hilfe der beobachteten Werte von  $X_1$  und der beobachteten sowie unbeobachteten, aber mit Platzhaltern gefüllten Werte der anderen Variablen ( $X_2 \dots X_8$ ) vorhergesagt. Die vorhergesagten Werte (#) werden dann anstelle des Platzhalters in  $X_1$  imputiert und man erhält eine Variable mit imputierten Werten (hier genannt:  $X_{1+t}$ ). Das Ergebnis dieser Schätzung wird in der folgenden Tabelle A5 veranschaulicht.

Tabelle A5: Datensatz mit geschätzten Werten in  $X_1$  und beobachteten Werten sowie Platzhaltern in  $X_2 \dots X_8$

	$X_{1+t}$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$
1	+	+	*	+	+	+	*	*
...	#	+	+	+	+	*	+	+
n	#	+	+	+	+	*	+	+

Der FCS-Algorithmus fährt dann mit der Schätzung der fehlenden Werte von  $X_2$  fort, indem der Platzhalter für diese Variable gelöscht und erneut ein eigens für  $X_2$  spezifiziertes Imputationsmodell zur Schätzung benutzt wird. Die Logik dieser Schätzung wird hier wiederum mit der folgenden, vereinfachten Gleichung (ohne Gewichtungen und ohne Zufallskomponente) veranschaulicht:

$$X_{2,\text{mis}} = X_{2,\text{obs}} + X_3 \dots X_8 + X_{1+t}$$

Wie mit dieser vereinfachten Gleichung leicht zu erkennen ist, werden auch hier (wie im ersten Schritt) alle verfügbaren Informationen herangezogen, um die fehlenden Werte in  $X_2$  zu ersetzen. Das Besondere an diesem Vorgehen ist, dass die FCS-Methode in jedem ihrer Schritte alle Informationen verwendet, die aus den vorangegangenen Schritten gewonnen werden können (in diesem Bsp. ist das  $X_{1+t}$ ). Dieser Vorgang kann als „machine learning“ bezeichnet werden: der FCS-Algorithmus gewinnt in einem ersten Schritt bestimmte Informationen und verwendet diese, um weitere Informationen zu gewinnen, welche wiederum dazu benutzt werden, um neue Informationen zu schaffen.

Eine Iteration nach der FCS-Methode dauert in unserem Beispiel so lange, bis die fehlenden Werte in  $X_8$  nach folgender Gleichung (wiederum stark vereinfacht dargestellt, s.o.) ersetzt sind:

$$X_{8,\text{mis}} = X_{8,\text{obs}} + X_{1+t} \dots X_{7+t}$$

Ist das der Fall, liegt nach dem ersten Iterationsschritt ein vollständiger Datensatz vor, der entsprechend unseres Beispiels mit der folgenden Tabelle A6 veranschaulicht werden kann.

Tabelle A6: Datensatz mit vollständig geschätzten, fehlenden Werten (#) in allen Variablen

	X1+t	X2+t	X3+t	X4+t	X5+t	X6+t	X7+t	X8+t
1	+	+	#	+	+	+	#	#
...	#	+	+	+	+	#	+	+
n	#	+	+	+	+	#	+	+

Der *zweite* Iterationsschritt beginnt mit der erneuten Ersetzung der Missingwerte in X<sub>1</sub>. Dabei werden aus X<sub>1</sub> die imputierten Missings entfernt, sodass dann wiederum die ursprüngliche Variable mit ihren fehlenden Werten vorliegt, die im zweiten Iterationsschritt erneut geschätzt werden müssen (im Folgenden veranschaulicht durch die Tabelle A7).

Tabelle A7: Datensatz zur Schätzung der Missings (-) in X<sub>1</sub> im zweiten Iterationsschritt

	X <sub>1</sub>	X <sub>2+t</sub>	X <sub>3+t</sub>	X <sub>4+t</sub>	X <sub>5+t</sub>	X <sub>6+t</sub>	X <sub>7+t</sub>	X <sub>8+t</sub>
1	+	+	#	+	+	+	#	#
...	-	+	+	+	+	#	+	+
n	-	+	+	+	+	#	+	+

Im Unterschied zum ersten Iterationsschritt werden im zweiten Iterationsschritt die Missings nun mit den ersetzten fehlenden Werten in X<sub>2</sub> bis X<sub>8</sub> (X<sub>2+t</sub> ... X<sub>8+t</sub>) vorhergesagt und keine Platzhalter benutzt, was wiederum (stark vereinfacht, s.o.) mit der folgenden Gleichung verdeutlicht werden kann:

$$X_{1,\text{mis}} = X_{1,\text{obs}} + X_{2+t} \dots X_{8+t}$$

Im zweiten Iterationsschritt erfolgt sodann die Ersetzung der Missings in X<sub>2</sub> unter Verwendung der ersetzten Daten in X<sub>1</sub> (genannt: X<sub>1+t2</sub>), was wiederum (stark vereinfacht, s.o.) mit der folgenden Gleichung verdeutlicht werden kann:

$$X_{2,\text{mis}} = X_{2,\text{obs}} + X_{3+t} \dots X_{8+t} + X_{1+t2}$$

Wenn der FCS-Algorithmus die Schätzungen der fehlenden Werte von Variable X<sub>1</sub> bis Variable X<sub>8</sub> so häufig wiederholt hat, wie es der Anzahl der vorgegebenen Iterationen entspricht (als Voreinstellung in SPSS: 10mal), liegt eine vollständige Iterationskette und damit ein vollständig imputierter Datensatz vor. Für jeden weiteren, im MI-Verfahren zu imputierenden Datensatz (m) startet SPSS wieder mit der Ziehung der Platzhalter (s.o.) und beginnt die Iterationskette wieder von vorne.

Wenn die im MI-Verfahren gewünschte Zahl an imputierten Datensätzen (m) vorliegt, ist mit der Analysephase des MI-Verfahrens fortzufahren (vgl. SISS-Kapitel 3.3).



## Literaturverzeichnis

- Acock, A.C., 2005: Working with Missing Values. *Journal of Marriage and Family* 67: 1012-1028.
- Allison, P.D., 2000: Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research* 28: 301-309.
- Allison, P.D., 2002: *Missing Data*. Thousand Oaks: Sage.
- Allison, P.D., 2015: Imputation by Predictive Mean Matching: Promise & Peril. <http://statisticalhorizons.com/predictive-mean-matching> (zugegriffen am 2.1.2017).
- Arbuckle, J.L., 1996: Full Information Estimation in the Presence of Incomplete Data. S. 243-277 in: Marcoulides, G.A./ Schumacker, R.E. (eds.), *Advanced Structural Equation Modeling*. Mahwah: Lawrence Erlbaum.
- Baltes-Götz, B., 2013: *Behandlung fehlender Werte in SPSS und Amos*. Trier: Universität Trier.
- Carpenter, J.R./ Kenward, M.G., 2013: *Multiple Imputation and its Application*. New York: Wiley.
- Cohen, J./ Cohen, P./ West, S.G./ Aiken, L.S., 2003: *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences* (3rd edition). Hillsdale: Erlbaum.
- Cole, J.C., 2008: How to Deal with Missing Data. Conceptual Overview and Details for Implementing Two Modern Methods. S. 214-238 in: Osborne, J.W. (ed.), *Best Practices in Quantitative Methods*. Thousand Oaks: Sage.
- Enders, C., 2010: *Applied Missing Data Analysis*. New York: Guilford.
- Field, A., 2013: *Discovering Statistics Using SPSS* (4th edition). London: Sage.
- Fox, J., 2016: *Applied Regression Analysis and Generalized Linear Models* (3rd edition). Los Angeles: Sage.
- Gabler, S./ Ganninger, M., 2010: Gewichtung. S. 143-164 in: Wolf, C./ Best, H. (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse*. Wiesbaden: VS Verlag.
- Graham, J.W., 2009: Missing Data Analysis: Making it Work in the Real World. *Annual Review of Psychology* 60: 549-576.
- Graham, J.W./ Hofer, S.M., 2000: Multiple Imputation in Multivariate Research. S. 201-218 in: Little, T.D. et al. (eds.), *Modeling Longitudinal and Multilevel Data*. Mahwah, New Jersey: Lawrence Erlbaum.
- Graham, J.W./ Schafer, J.L., 1999: On the Performance of Multiple Imputation for Multivariate Data With Small Sample Size. S. 1-23 in: Hoyle, R.H. (ed.), *Statistical Strategies for Small Sample Research*. London: Sage.
- Li, C., 2013: Little's Test of Missing Completely at Random. *The Stata Journal* 13: 795-809.
- Little, R.J.A., 1988: A Test of Missing Completely at Random for Multivariate Data with Missing Value. *Journal of the American Statistical Association* 83: 1198-1202.
- Little, R.J.A./ Rubin, D.B., 2002: *Statistical Analysis with Missing Data* (2nd edition). Hoboken, N.J.: Wiley.
- Osborne, J.W., 2013: Dealing with Missing or Incomplete Data. S. 105-138 in: Osborne J.W., *Best Practices in Data Cleaning*. London: Sage.
- Pigott, T.E., 2001: A Review of Methods for Missing Data. *Educational Research and Evaluation* 7: 353-383.

- Schafer, J.L., 1997: Analysis of Incomplete Multivariate Data. New York: Taylor & Francis.
- Schafer, J.L./ Graham, J., 2002: Missing Data: Our View of the State of the Art. Psychological Methods 7: 147-177.
- Schafer, J.L./ Olsen, M.K., 1998: Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research 33: 545-572.
- Scheffer, J., 2002: Dealing with Missing Data. Research Letters in the Information and Mathematical Sciences 3: 153-160.
- Schulz, S., 2015: Gewichtung von ALLBUS-Umfragedaten – ALLBUS 2014. S. iii-iv in: Baumann, H./ Sonja, S., ALLBUS 2014 – Variable Report. Köln; GESIS.
- Tabachnick, B.G./ Fidell, L.S., 2013: Using Multivariate Statistics (6th edition). Boston: Allyn & Bacon.
- Urban, D., 1993: Logit-Analyse: Statistische Verfahren zur Analyse von Modellen mit qualitativen Response-Variablen. Stuttgart: Fischer.
- Urban, D./ Mayerl, J., 2011: Regressionsanalyse: Theorie, Technik und Anwendung (4. Auflage). Wiesbaden: Springer.
- Urban, D./ Mayerl, J., 2014: Strukturgleichungsmodellierung. Ein Ratgeber für die Praxis. Wiesbaden: Springer.
- Van Buuren, S., 2007: Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. Statistical Methods in Medical Research 16: 219-242.
- Von Hippel, P.T., 2009: How to Impute Squares, Interactions, and Other Transformed Variables. Sociological Methodology 30: 265-291.

**SISS:****Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart**

<http://www.uni-stuttgart.de/soz/institut/forschung/siss.html>

- No. 44/2016      Regressionsanalyse bei fehlenden Variablenwerten (missing values): Imputation oder Nicht-Imputation? Eine Anleitung für die Regressionspraxis mit SPSS.  
Dieter Urban, Jochen Mayerl, Andreas Wahl
- No. 43/2014      Meta-Analyse mit Mehrebenenmodellen unter Verwendung von HLM.  
Joachim Fiebig, Dieter Urban
- No. 42/2013      Panelanalyse mit Mehrebenenmodellen. Eine anwendungsorientierte Einführung.  
Thomas Krause, Dieter Urban
- No. 41/2012      Meta-Analyse und "publication bias". Eine empirisch-statistische Fallstudie.  
Joachim Fiebig, Gerrit Rüdebusch, Dieter Urban
- No. 40/2011      Pädosexuelle Viktimisierung und pädosexuelle Straffälligkeit. Ein „harter“ empirischer Hypothesentest.  
Dieter Urban, Joachim Fiebig
- No. 39/2010      Binär-logistische Regressionsanalyse.  
Grundlagen und Anwendung für Sozialwissenschaftler.  
Jochen Mayerl, Dieter Urban
- No. 38/2010      Das Working-Poor-Problem in Deutschland.  
Empirische Analysen zu den Ursachen von Armut trotz Erwerbstätigkeit.  
Leonie Hellmuth, Dieter Urban
- No. 37/2010      Der Bystander-Effekt in alltäglichen Hilfesituationen: Ein nicht-reaktives Feldexperiment.  
Katrin Alle, Jochen Mayerl
- No. 36/2008      Berufseintritt und Berufssituation von Soziologieabsolventen der Universität Stuttgart.  
Deskriptive Ergebnisse einer Absolventenbefragung aus dem Jahr 2007.  
Jochen Mayerl, Dieter Urban
- No. 35/2007      Antwortlatenzzeiten in TRA-Modellen. Zur statistischen Erklärung von (Geld)-Spendenverhalten.  
Dieter Urban, Jochen Mayerl
- No. 34/2006      Zur soziologischen Erklärung individuellen Geldspendens.  
Eine Anwendung und Erweiterung der Theory of Reasoned Action unter Verwendung von Antwortlatenzzeiten in einem Mediator-Moderator-Design.  
Jochen Mayerl
- No. 33/2006      Ist Bildung gleich Bildung? Der Einfluss von Schulbildung auf ausländerablehnende Einstellungen in verschiedenen Alterskohorten.  
Klaus Hadwiger
- No. 32/2005      Analyzing cognitive processes in CATI-Surveys with response latencies: An empirical evaluation of the consequences of using different baseline speed measures.  
Jochen Mayerl, Piet Sellke, Dieter Urban
- No. 31/2005      Persönliche Netzwerke in der Systemtheorie.  
Jan A. Fuhse
- No. 30/2004      Erhöht ein Opfer-Täter-Zyklus das Risiko, Sexualstraftaten als pädosexuelle Straftaten zu begehen? Ergebnisse einer ereignisanalytischen Pilotstudie.  
Dieter Urban, Heiko Lindhorst

**SISS:****Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart**

<http://www.uni-stuttgart.de/soz/institut/forschung/siss.html>

- No. 29/2003 Können Nonattitudes durch die Messung von Antwortreaktionszeiten ermittelt werden? Eine empirische Analyse computergestützter Telefoninterviews.  
Jochen Mayerl
- No. 28/2003 Systeme, Netzwerke, Identitäten. Die Konstitution sozialer Grenzziehungen am Beispiel amerikanischer Straßengangs.  
Jan A. Fuhse
- No. 27/2002 Probleme bei der Messung individueller Veränderungsraten. 13 empirisch und methodisch induzierte Effekte, die es schwierig machen, Veränderungen von generalisierten Bewertungen zu ermitteln.  
Dieter Urban
- No. 26/2002 Vertrauen und Risikoakzeptanz. Zur Relevanz von Vertrauen bei der Bewertung neuer Technologien.  
Martin Slaby, Dieter Urban
- No. 25/2002 Risikoakzeptanz als individuelle Entscheidung. Zur Integration der Risikoanalyse in die nutzentheoretische Entscheidungs- und Einstellungsforschung.  
Martin Slaby, Dieter Urban
- No. 24/2001 Ist Glück ein affektiver Sozialindikator subjektiven Wohlbefindens? Dimensionen des subjektiven Wohlbefindens und die Differenz zwischen Glück und Zufriedenheit.  
Jochen Mayerl
- No. 23/2001 Religiosität und Profession. Longitudinale Analysen zur Entwicklung des religiösen Selbstbildes bei Erzieherinnen.  
Heiko Lindhorst
- No. 22/2001 Differentielle Technikakzeptanz, oder: Nicht immer führt die Ablehnung einer Technik auch zur Ablehnung ihrer Anwendungen. Eine nutzentheoretische und modell-statistische Analyse.  
Martin Slaby, Dieter Urban
- No. 21/2001 Unser „wir“ - ein systemtheoretisches Modell von Gruppenidentitäten.  
Jan A. Fuhse
- No. 20/2000 Längsschnittanalysen mit latenten Wachstumskurvenmodellen in der politischen Sozialisationsforschung.  
Dieter Urban
- No. 19/2000 Die Zufriedenheit von Stuttgarter Studierenden mit ihrer Lebens- und Wohnsituation. Erste deskriptive Ergebnisse einer sozialwissenschaftlichen Studie zu allgemeinen und bereichsspezifischen Zufriedenheiten der Studierenden des Campus Vaihingen und des Campus Hohenheim.  
Projektgruppe Campus: Slaby, M.; Grund, R.; Mayerl, J.; Noak, T.; Payk, B.; Sellke, P.; Urban, D.; Zudrell, I.
- No. 18/1999 Role Models and Trust in Socio-Political Institutions: A Case Study in Eastern Germany, 1992-96.  
Joachim Singelmann, Toby A. Ten Ayck, Dieter Urban
- No. 17/1998 Zur Interaktion zwischen Befragten und Erhebungsinstrument. Eine Untersuchung zur Konstanz des Meinungsurteils von Befragten im Interviewverlauf.  
Martin Slaby

**SISS:****Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart**

<http://www.uni-stuttgart.de/soz/institut/forschung/siss.html>

- No. 16/1998     Technikeinstellungen: gibt es die überhaupt? Ergebnisse einer Längsschnittanalyse von Bewertungen der Gentechnik.  
Dieter Urban
- No. 15/1998     Bewertende Einstellungen zur Gentechnik: ihre Form, ihre Inhalte und ihre Dynamik. Kurzbericht zu Ergebnissen des Forschungsprojektes „Einstellungen zur Gentechnik“.  
Dieter Urban, Uwe Pfenning, Joachim Allhoff
- No. 14/1997     Politische Sozialisation im Transformationsprozeß: Die Entwicklung demokratiebezogener Einstellungen von ostdeutschen Jugendlichen und deren Eltern zwischen 1992 und 1996.  
Barbara Schmidt, Dieter Urban, Joachim Singelmann
- No. 13/1997     Die soziale Vererbung von „Ausländerfeindlichkeit“. Eine empirische Längsschnittanalyse der intra- und intergenerativen Transmission von sozialen Einstellungen.  
Dieter Urban und Joachim Singelmann
- No. 12/1997     Gentechnik: „Fluch oder Segen“ versus „Fluch und Segen“. Bilanzierende und differenzierende Bewertungen der Gentechnik in der öffentlichen Meinung.  
Dieter Urban und Uwe Pfenning
- No. 11/1996     Einstellungsmessung oder Einstellungsgenerierung? Die Bedeutung der informationellen Basis bei Befragten für die empirische Rekonstruktion von Einstellungen zu gentechnischen Anwendungen.  
Martin Slaby
- No. 10/1996     Jugend und Politik im Transformationsprozeß. Eine Fallstudie zur Stabilität und Veränderung von politischen Einstellungen bei ostdeutschen Jugendlichen zwischen 1992 und 1995.  
Dieter Urban, Joachim Singelmann, Helmut Schröder
- No. 09/1995     Social Indicators in a Nonmetropolitan County: Testing the Representativeness of a Regional Nonrandom Survey in Eastern Germany.  
Dieter Urban, Joachim Singelmann
- No. 08/1995     Handbuch zur empirischen Erhebung von Einstellungen/Kognitionen zur Bio- und Gentechnologie (inklusive Diskette)  
(zweite, überarbeitete und erweiterte Auflage)  
Uwe Pfenning, Dieter Urban, Volker Weiss
- No. 07/1995     Latent Attitude Structures Directing the Perception of New Technologies. An Application of SEM-Methodology to the Construction of Attitude Measurement Models Related to Technologies of Prenatal Genetic Engineering and Testing.  
Dieter Urban
- No. 06/1995     Regionale Strukturprobleme.  
Sammelband der Beiträge zum Symposium vom 13. und 14. Oktober 1994.  
Frank C. Englmann (Hrsg.)
- No. 05/1995     Die Bedeutung Neuronaler Netze in der Ökonomie.  
Hermann Schnabl
- No. 04/1994     Rechtstatsachen zum Unternehmens- und Gesellschaftsrecht aus den neuen Bundesländern.  
Udo Kornblum

**SISS:****Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart**

<http://www.uni-stuttgart.de/soz/institut/forschung/siss.html>

- No. 03/1994    Neue württembergische Rechtstatsachen zum Unternehmens- und Gesellschaftsrecht.  
Udo Kornblum
- No. 02/1994    Report on the German Machine Tool Industry.  
Frank C. Englmann, Christian Heyd, Daniel Köstler, Peter Paustian  
with the assistance of Susanne Baur and Peter Bergmann
- No. 01/1994    "Vertrauen" - soziologisch betrachtet. Ein Beitrag zur Analyse binärer Interaktionssysteme.  
Peter Antfang, Dieter Urban